UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Applying Generalized Linear Models to Estimate Group Size and Improve Blainville's Beaked Whale Abundance Estimation

Patrícia Alexandra de Almeida Jorge

Mestrado em Bioestatística

Trabalho de projeto orientado por:
Doutor Tiago Marques
Doutora Helena Mouriño

2017

Fairy tales are more than true: not because they tell us that dragons exist, but because they tell us that dragons can be beaten.

- Neil Gaiman

Acknowledgements

This work was conducted over data collected under the project GROUPAM, funded by the Office of Naval Research of the United States of America. I thank Jessica Ward, Karin Dolan, David Moretti, and Len Thomas, as well as the remaining large NUWC team, which were responsible for processing the raw acoustic data into a format that I could use, and for answering some of the questions that arose while implementing the analysis. An additional thank you to Karin Dolan for having gone through the hoops and loops of getting the data through the public release process.

I would also like to thank Helena Mouriño, for her formidable knowledge, support and patience.

My sincere gratitude to Tiago Marques, who made all this possible. I could not have had someone better to guide me through this journey.

Table of Contents

List of Figures

List of Tables

Resumo

Em ecologia, métodos precisos e eficientes são fundamentais no que toca à estimação da abundância das populações naturais, sendo necessários para uma gestão e conservação sustentáveis e eficazes. Consequentemente, é importante optimizar os modelos existentes de maneira a garantir a eficácia das suas previsões. Assim, métodos que garantam a monitorização com a mínima intervenção humana têm vindo a ganhar popularidade no estudo das populações naturais.

A Baleia de bico de Blainville (*Mesoplodon densirostris*, *Md*) é a espécie do género *Mesoplodon* com a mais abrangente área de distribuição, estando presente em águas temperadas e tropicais de todos os oceanos. São facilmente identificadas pelo seu corpo largo e robusto, bem como pelo bico bem definido que está na origem do seu nome: "*densirostris*" vem do Latim que significa "do bico denso". Apesar da sua ampla distribuição, raramente é avistada devido a passar a maior parte do tempo em águas a grandes profundidades. Apenas por breves períodos se desloca à superfície, o que resulta numa baixa probabilidade de avistamento. Esta espécie é também conhecida por se associar em grupos e exibir um comportamento metronómico aquando o mergulho, apresentando tendência para vocalizar apenas durante os mergulhos profundos que efectua para a alimentação. Para tal, recorre a sinais de ecolocalização ultrasónica de banda larga, conhecidos como "cliques", com um comprimento de onda de 26 a 51 kHz. Estes cliques ocorrem predominantemente em locais onde as baleias procuram e encontram presas, e são divididos em duas categorias: *search clicks* e *buzz clicks*. Os primeiros são produzidos durante todo o mergulho profundo, enquanto que os últimos são emitidos durante curtos períodos no estádio final da captura de presa. Devido à continuidade dos *search clicks* ao longo do mergulho, é neste primeiro tipo de cliques que o presente trabalho se baseia.

O AUTEC (Atlantic Undersea Test and Evaluation Center) é um centro naval de treino pertencente aos E.U.A localizado na Tongue of the Ocean (TOTO), nas Bahamas. É um local onde frequentemente ocorrem testes com sonares, e onde a espécie da baleia em estudo é rotinamente detetada. O centro possui um vasto campo de 93 hidrofones ligados à base, aos quais se pode recorrer para detetar os cliques de ecolocalização produzidos por *Md*. Dadas as características dos hidrofones, a densa grelha que estes apresentam e a área que ocupam, combinando com características intrínsecas à espécie, os mergulhos efetuados dentro do campo considerado serão certamente detetados.

Recorrendo a dados recolhidos pelo AUTEC, é apresentado um método para estimar a abundância de *Md*. Visto os indivíduos desta espécie despenderem muito pouco tempo à superfície, os tradicionais métodos de estimação de abundância, como a amostragem por distâncias por transetos lineares, podem conduzir a resultados inconclusivos. Com o auxílio de métodos acústicos que detetam e classificam os cliques de ecolocalização de *Md*, é possível neste trabalho atribuir estes cliques detetados a cada grupo que efectue um mergulho.

A abordagem proposta propõe desenvolver métodos previamente apresentados por DiMarzio *et al.* (2008) e Moretti *et al.* (2010). De acordo com estes autores, a densidade de animais é estimada como o produto entre uma estimativa da densidade de grupos e do número médio de animais por grupo. Ao invés de se considerar um tamanho médio de grupo baseado na literatura,

o número de animais em cada grupo será estimado com base na sua pegada acústica (*acoustical footprint*) através de um modelo linear generalizado.

Para este estudo, são considerados dois conjuntos de dados:

(1) o conjunto de dados da modelação. Utilizado para construir o modelo do tamanho de grupo como função da pegada acústica dos grupos;

(2) o conjunto de dados da estimação da densidade. Utilizado para estimar a densidade dado o modelo de tamanho de grupos obtido com o primeiro conjunto de dados.

O conjunto de dados da modelação consiste em 51 mergulhos profundos identificados entre 2005 e 2008, para os quais o tamanho de grupos se confirmou visualmente ou mediante uma análise dos dados acústicos complexa e que como tal não pode ser automatizada ou rotineiramente utilizada. As potenciais variáveis explicativas incluem, para cada mergulho detetado, o número de hidrofones envolvidos na deteção dos cliques, o número de cliques detectados em cada hidrofone, o início e o fim do período da ecolocalização, o tempo entre cliques sucessivos detectados num mesmo hidrophone, e variáveis binárias: uma que indica se algum dos hidrofones em que o grupo foi detectado se encontra na periferia da rede de hidrofones, e outras duas que indicam se algum dos hidrofones pertencem a uma categoria diferente dos restantes (hidrofones "*Whiskey* ou "*Direction*"). Posteriormente, construíram-se variáveis adicionais tendo por base estas últimas: a duração dos cliques e a taxa a que os cliques ocorrem. O tamanho de grupo para estes dados varia entre 1 e 6 baleias.

O conjunto de dados da estimação da densidade é uma série temporal que cobre cerca de 4 meses do ano de 2011: (1) de 28 de April a 27 de Junho (61 dias), (2) de 20 de Outubro a 6 de Novembro (18 dias), e (3) de 2 a 31 de Dezembro (30 dias). Estes dados foram processados utilizando o mesmo procedimento que gerou os dados para a modelação do tamanho de grupo, sendo que este será estimado para todos os mergulhos profundos detetados. Este método permite quantificar os cliques que ocorreram no campo de hidrofones do AUTEC durante o período considerado, permitindo também estimar o número total de animais envolvidos. Por sua vez, tal procedimento permitirá a estimação de densidade ao longo do tempo recorrendo a um método melhorado de contagem de mergulhos proposto por Moretti *et al* (2010).

Num total de 15493 potenciais mergulhos apenas 8271 foram considerados após implementar um pré-processamento dos dados baseado em características biológicas de *Md* e dos hidrofones. Este pré-processamento consiste em excluir possíveis falsos positivos tendo em conta:

(1) deteções que ocorrem somente num único hidrofone;

(2) um mínimo de 400 cliques detetados por grupo;

(3) grupos apenas detetados por hidrofones localizados na periferia.

As 8271 deteções consideradas como verdadeiros mergulhos das baleias aparentam dividir-se de forma uniforme ao longo dos 3 períodos considerados: 4562 para o primeiro, com uma média de 75 mergulhos detetados por dia; 1439 para o segundo, com cerca de 80 mergulhos por dia; e 2270 mergulhos para o terceiro período, com uma média de 76 mergulhos por dia. A análise indica que o tamanho de grupo poderá ser previsto pela pegada acústica do grupo com base nas covariáveis consideradas. A variável mais importante para a modelação do tamanho do grupo aparenta ser a taxa de cliques. No entanto, será necessária uma maior recolha de dados de modelação para sustentar esta hipótese.

Aquando da estimação, verifica-se que existe uma certa flutuação da densidade ao longo do tempo. De maneira a propagar a variância do modelo selecionado pelas estimativas de

variância da densidade por dia, implementou-se um bootstrap. Tal conduziu a novas estimativas dos parâmetros, o que por sua vez faculta diferentes estimações para cada tamanho de grupo. Este procedimento permite visualizar possíveis variações na estimação dos parâmetros e a sua influência na estimação do tamanho de grupo e da densidade para cada dia.

De futuro, pretende-se relacionar esta flutuação com a ocorrência de fatores externos, nomeadamente fatores antropogénicos. Os resultados deste trabalho, conjugando com trabalhos anteriores e futuros, serão utilizados para prever os comportamentos desta espécie de maneira a monitorizar padrões inerentes à sua mobilidade. Com isto, espera-se contribuir para o repositório de informação de *Md* e preencher lacunas na compreensão dos hábitos desta espécie.

Palavras-Chave: Baleia de bico de Blainville, Ecolocalização, Contagem de mergulhos, Estimação de densidade, Tamanho de grupo, Acústica passiva.

Abstract

Blainville's beaked whales (*Mesoplodon densirostris*, *Md*) are known to associate in groups, exhibiting metronomic dive behaviour. They tend to vocalize via echolocations only during deep foraging dives using broadband clicks.

Using *Md* click data collected on AUTEC (Atlantic Undersea Test and Evaluation Center) hydrophones, a method for estimating *Md* abundance is presented. The *Md* click data accounts for the echolocations for each corresponding *Md* group foraging dive, where the start of a foraging dive is assumed to be the time of the first detected echolocation click.

The proposed approach extends previous methods developed by Moretti *et al.* (2010) and DiMarzio *et al.* (2008). Instead of considering an estimated average group size value based on literature, the size of each group will be estimated considering variables derived from the acoustic data, via a generalized linear model.

We consider two different data sets: one to build the model of group size as a function of the groups acoustic footprint, and another to estimate density, leveraging on the group's size model.

The modelling dataset consists of 51 deep dives identified between 2005 and 2008, for which the group size was visually confirmed. Potential explanatory variables include, for each detected dive, the number of the hydrophones which detected the echolocation clicks, the number of clicks detected in each hydrophone, the corresponding start and end of the echolocation period, and binary variables which indicate whether or not the particular group had its clicks detected by at least one hydrophone located on the edge, or if at least one hydrophone belongs to the particular types of Whiskey or Bi-directional hydrophones. Further, a number of derived variables were constructed from the dataset. The group size in this modelling data ranged between 1 and 6 whales.

The density estimation dataset is a time series of AUTEC data from which density will be estimated. It includes 3 separate periods of time in 2011: (1) 61 days from the 28th of April to the 27th of June, (2) 18 days from the 20th of October to the 6th of November, and (3) 30 days from the 2nd to the 31st of December. These data were processed using the same procedure that generated the data for the group size model, and the group size will be estimated for all deep dives detected. This method allows to quantify how many dives occurred on the AUTEC range during that period, and to estimate the total number of animals involved. This in turn allows the estimation of density over time using this improved version of the dive counting method proposed by Moretti *et al.* (2010).

In total 15493 potential deep dives were detected in the second dataset. A preprocessing of the data to exclude false positives was implemented, based on a set of biologically infeasible characteristics:

(1) detections occurring on a single hydrophone;

(2) a minimum threshold of 400 clicks detected. This resulted in a much more biologically plausible distribution of observed vocal lengths, matching what would be expected given described values in the literature;

(3) groups detected only on edge hydrophones, considering these would correspond to groups outside the area of inference.

This led to 8271 detections considered to correspond to relevant beaked whale deep dives. The first period of time recorded 4562 dives, with an average of 75 dives per day; the second period showed 1439 dives with an average of 80 dives per day; and the third one registered 2270 dives with an average of 76 dives per day.

After adjusting generalized linear models, there is an indication that the group size can be predicted from acoustic footprint of the group via available covariates. The most important variable to explain group size appears to be the click rate. When looking at the estimation's results, a certain fluctuation over time is noticeable. Hereafter, this fluctuation is intended to be related with external factors, namely anthropogenic factors. A bootstrap was then applied to propagate the variance in the model of group size thorough the estimates of variance of density per day.

The results from this study, conjugating with previous and future studies, will allow a better understanding of this species behaviour in order to monitor mobility patterns.

Keywords: Blainville's beaked whales, Echolocation, Dive counting, Density Estimation, Group size, Passive acoustic.

Chapter 1

Introduction

1.1 The Idea Behind this Study

Efficient and precise methods for estimating the abundance of natural populations are required for their effective management and conservation. Consequently, it is important to optimize existing models. Methods allowing *in situ* monitoring with a minimum amount of human intervention are becoming more popular to study natural populations. As the individuals belonging to the *Mesoplodon densirostris* species spend little time on the surface, other traditional abundance estimation methods, like line transect distance sampling, may lead to inconclusive results. The fact that these whales produce distinctive echolocation clicks at a relatively steady rate while searching for prey, makes them a suitable candidate for Passive Acoustic Monitoring (PAM) (Tyack et al., 2006).

Group size is an important factor to account for when dealing with animal density estimation. Describing the group size distribution over time and space might bring further knowledge about the effect of AUTEC sonar usage on the considered species (Marques *et al.*, 2013); and it was in fact shown by DiMarzio *et al.* (2008) that, for a reduced number of groups with known size, the acoustical footprint of a group is dependent on its group size.

The present case study focuses on PAM to detect and classify these whales' echolocation clicks. It starts by using a first dataset to model group size as a function of the acoustical footprint of the groups on the surrounding hydrophones on the Atlantic Undersea Test and Evaluation Center. A model is created relating acoustic footprint statistics (e.g., click detection counts, number of hydrophones involved) on hydrophones to group size, estimating the parameters using surface visual observations. The statistical model will enable the development of a real-time algorithm to estimate and display group size information for support of routine density estimation (e.g., Marques *et al.*, 2009 and Moretti *et al.*, 2010) and to assist in live range operations.

A previously published approach to estimate this species density in the area uses an estimated average group size value based on literature (Moretti *et al*, 2010). As visually confirming each group's size for time-series data is an impossible task, an automated way to estimate group size is needed. This work will extend the previous approach by first modelling the group size resorting to a dataset of 51 *Md* groups detected at the hydrophones on the Atlantic Undersea Test and Evaluation Center whose size was visually and acoustically confirmed, relating the group size to the acoustic footprint variables.

Resorting to the previous developed model, group size is estimated for more than 8000 *Mesoplodon Densirostris* groups, whose group size was not visually confirmed, from a dataset where density will be estimated: a time series of acoustic data also collected by the same hydrophones, which considers the same acoustic footprint variables from the modelling dataset. After estimating group size for each detected group, density and associated precision measures

will be estimated by day over the time period for which recordings are available.

In the next subsections, a brief description of *Mesoplodon Densirostris* will take place, as well as the hydrophone range camp, followed by a summary of the main goals.

1.2 Blainville's Beaked Whale

Description

Blainville's beaked whale, *Mesoplodon densirostris* (*Md*), is the widest ranging species belonging to the genus *Mesoplodon*, occurring in low to mid-latitudes in all oceans. Maximum recorded length is around 4.7 meters, with the individuals weighing about 1000 kilograms (Jefferson *et al.*, 2008). They are most easily identified by their large and robust body, small forehead and long, dense, well-defined beak, which inspired this species' name. The most distinctive feature is the dense upper jawbone, along with the posterior half of the lower jaw highly arched with two massive horn-like teeth, typically more prominent in males. Males also tend to show dorsal body scarring, whose patterns seem to match the tooth structure and position of conspecifics, which suggests these same markings occur due to intraspecific combat (MacLeod, 1998). Body colouration is lighter on female individuals, varying from blue-grey to black, with a whiter tone on the ventral side and several spots along the body. (Leatherwood & Reeves, 1983; McCann, 1963; Mead, 1989; Pastene *et al.*, 1990). Figure 1.1 illustrates these *Md*'s physical features.

Figure 1.1: Representation of Blainville's beaked whales. Female (bottom) and male (top). The later exhibits body scarring, more prominent teeth and darker body colouration. Source: © Wurtz – www.artescienza.org.

Behaviour and Habits

Although it is perhaps one of the most well documented beaked whale species, these whales exhibit a shy and discreet behaviour, being rarely seen due to spending most of their time foraging at depth. Only a short period of time is spent at the surface, resulting in a low probability of visual detection (Barlow, 1999; Tyack *et al.*, 2006). These whales typically occur

in small groups of up to about 11 individuals (Dimarzio *et al*, 2008), and engage in prolonged dives several times a day to feed mainly on squid; although they also prey on small deep-sea fish and crustaceans (Baird *et al*, 2008; Johnson *et al*, 2006). When foraging, the group is known to perform synchronized dives to great depths, time at which *Md* produces distinctive ultrasonic echolocation signals, known as 'clicks', with a bandwidth from 26 to 51 kHz (Johnson *et al*, 2006). Figure 1.2 provides an insight of *Md*'s diving profile, where the tagged individual forages to depths up to 1400 meters.

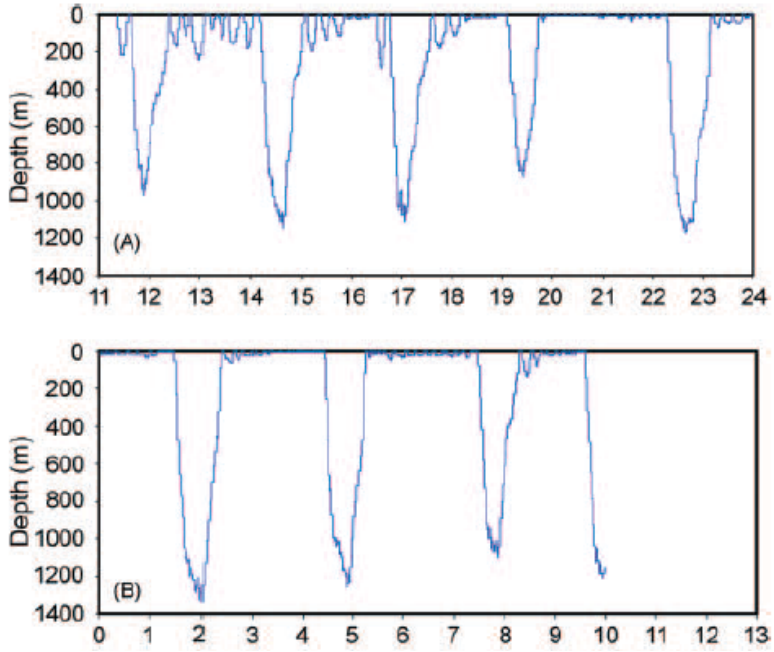


Figure 1.2: Illustration of an individual's diving profile depth: data from a 22.6h deployment on a tagged Blainville's beaked whale divided into two days: **A** (12.6 hours) and **B** (10 hours). Adapted from Baird *et al.*, 2006.

According to Johnson *et al.* (2006), there are two types of click sounds: search clicks (also known as 'regular clicks') and buzz clicks, as illustrated in figure 1.3. The first one is produced during the whole foraging dive, whereas the later is emitted in short bursts during the final stage of prey capture.

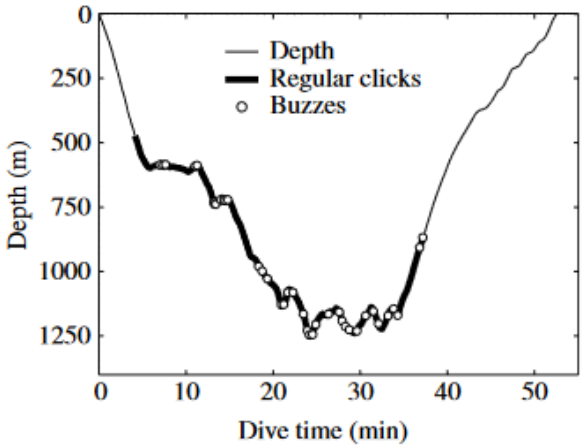


Figure 1.3: Dive profile illustration of a Blainville's beaked whale foraging dive showing vocal events, featuring regular and buzz clicks. Retrieved from Johnson *et al.*, 2006.

Tyack *et al.* (2006) suggests that *Md* hunt by echolocation in deep waters between 250 and 1900 meters, attempting to capture about 30 prey per dive. The food source is so deep that the average foraging dives are deeper (835 m) and longer (47 min) than reported in the literature for any other marine mammal species.

It is known that *Md* groups produce a minimum amount of clicks when diving (Moretti *et al.*, 2010, Shaffer *et al.*, 2013). According to Shaffer *et al.* (2013), the mean number of foraging clicks emitted by each animal is around 3000 clicks per dive (with a range of 939 – 6663 clicks).

Baird *et al.* (2008) reported that deep foraging dives (>800 m) occur at similar rates during both the day and night, despite whales spending more time in shallow depths (<100 m) during the night. Dives to mid-water depths (100-600 m) occurred significantly more often during the day. This suggests that the whales may spend less time in surface waters during the day to avoid near-surface, visually oriented predators such as large sharks or killer whales.

1.3 The Case Study

1.3.1 The AUTEC

The Atlantic Undersea Test and Evaluation Center (AUTEC) is a U.S.A. Navy testing and training range located in the Tongue of the Ocean (TOTO) in the Bahamas. It is a site of repeated sonar use, and *Md* are routinely detected year-round on the AUTEC range. It includes a large network of hydrophones cabled to shore that can be used to detect *Md* echolocation clicks. Given the hydrophone spacing and sensitivity, combined with the animals' clicks source level, all the dives occurring on the AUTEC range can be assumed to be detected with certainty (Moretti *et al.*, 2010).

As represented in figure 1.4, the training range consists of two separate hydrophones systems: the two older Whiskey arrays (hydrophones 1-14, a total of 14) which were the first devices installed, and the newer Advanced Hydrophone Replacement Program (AHRP) array (hydrophones 15-93, a total of 79) which hold a more recent technology. The AHRP

array is itself composed by two different types of hydrophones: 16 bi-directional (transmit and receive) and 63 uni-directional (receive only) hydrophones. The Whiskey and AHRP arrays have different hydrophone features and shore processing hardware resulting in distinct *Md* detection characteristics, while the uni-directional and bi-directional hydrophones have different receiver beam patterns. The bi-directional hydrophones have more electronic noise adding to a greater chance of false positive detections. The AHRP bi-directional hydrophones include numbers 15, 20, 30, 41, 42, 45, 56, 58, 61, 69, 72, 75, 78, 88, 91, and 93 (Shaffer, J., personal communication).

The range area is defined as a "convex hull" with a 6.5 km buffer around the non-edge hydrophones. This area for dive counting is defined based on the assumption that groups occurring within it would have at least some clicks detected on non-edge phones, and groups occurring outside that area would not have any clicks detected on non-edge phones. This provides a straightforward operational rule to include/exclude dives from our dive counting procedure, as will be explained forward.

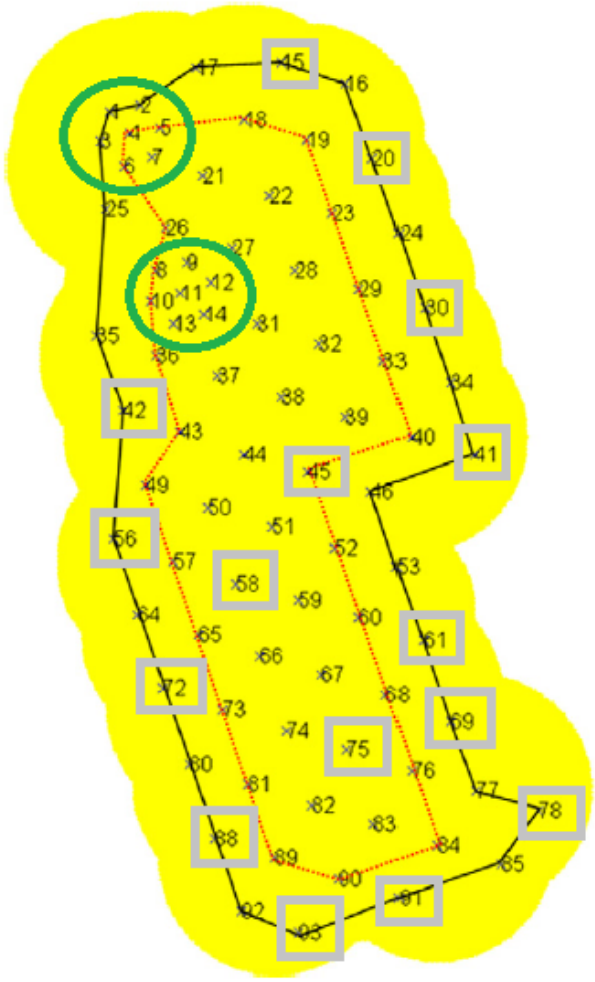


Figure 1.4: Hydrophone camp with the 93 hydrophones (numbered), featuring the "convex hull" area (yellow), the Edge hydrophones (black line), non-Edge hydrophones (red line and inside red line), Whiskey hydrophones (circled green), and Bidirectional hydrophones (squared grey). Adapted from Moretti *et al.*, 2010.

It is also important to highlight it is taken into account in this study whether or not a hydrophone is located on the edge. Since these type of hydrophones have a higher chance

of capturing echolocations out of the considered area of 1291km^2, it is more likely that they incorporate false positive detections. Edge hydrophones include numbers 1, 2, 3, 15, 16, 17, 20, 24, 25, 30, 34, 35, 41, 42, 46, 53, 56, 61, 64, 69, 72, 77, 78, 80, 85, 88, 91, 92 and 93.

1.4 Main Objectives

The specific key objectives of this study are:

1. To model group size as a function of the acoustic footprint of a group as detected automatically by an existing algorithm, using a dataset of acoustical footprints for groups with verified group size;
2. By using the model previously built, predict the group size for groups with no verified group size;
3. To estimate *Md* density per day for a 4 month period, adapting the previous method proposed by Moretti *et al.* (2010);
4. Obtain precision measures for the model predictions and density estimates using a non-parametric bootstrap.

Besides the four items above, two other "extra" objectives were incorporated in this project:

5. Since both datasets will probably be used in future studies, a more profound exploratory data analysis was performed, looking in particular at click detection differences between the hydrophone types.
6. Inspired in the author's own previous experience, this work's writing style is aimed at ecologists whose statistical knowledge might be scarse. Therefore, an extra effort was made to provide a theoretical explanation which may help the reader to actually understand what is behind the employed statistical methods.

In the next section, the methodology required to implement the methods will be described. Afterwards, the project's results are presented. We conclude with a discussion about our results, possible ways forward, and the summary of the acquired competencies during the MSc programme.

Chapter 2

Methodology

In this section we begin by describing the data, followed by all the transformations necessary to implement the analysis. All the data were analysed resorting to the *R* software.

2.1 The Data

For this study two different datasets are considered: (1) the modelling dataset consists of *Md* groups acoustic data whose size was visually confirmed. It was used to build the model of the group size as function of the group's acoustic footprint; and (2) the density estimation dataset which was employed after building the model, which consists on a time series of data from the AUTEC hydrophones for which the *Md* group size and density were estimated.

Both datasets were generated at AUTEC resorting to Autogrouper, which is a MATLAB (Mathworks) script, an automatic process that identifies whale dives by quickly identifying start and end times of the echolocations. It works by combining clicks within hydrophones into sequences of clicks, named click trains. Then it groups click trains close in space and time, i.e., detected simultaneously in adjacent hydrophones, into vocal groups. Each vocal group detected corresponds to a *Md* foraging dive. Associated with each detected dive there is a set of available statistics that define the acoustic footprint of the group, such as the amount of detected clicks and the quantity of hydrophones involved on each detection (Madsen *et al.*, 2013).

As buzz clicks are produced in short bursts with no FM structure and may be difficult to collect (Johnson *et al.*, 2006), the data only includes search clicks. Since these clicks are produced during the whole foraging dive, and not only during the final stage of prey capture (like buzz clicks), search clicks offer a thorough insight of *Md* acoustic footprint.

2.1.1 Modelling Dataset

The modelling dataset includes the Autogrouper routine output for 51 deep dives, between 2005 and 2008, that were confirmed either visually or resorting to a very detailed acoustical analysis which is far more time consuming than it would be possible to process data on an everyday basis. It includes several potential covariates to model group size:

- The number of hydrophones at which group i was detected, K_i;
- The number of clicks from group i detected at hydrophone k, $c_{i,k}$. It is then possible to obtain the total number of clicks detected for group i (N_i) by summing over the K_i hydrophones it was detected on:

$$N_i = \sum_{k=1}^{K_i} c_{i,k}\,; \tag{2.1}$$

- The mean number of clicks, m_i, detected per hydrophone for group i, where:

$$m_i = \frac{N_i}{K_i}\,; \tag{2.2}$$

- Each hydrophone's click period, in microseconds, from which it is possible to obtain the total clicking duration, d_i;
- The maximum click count per hydrophone detected for group i, $max_i(c_{i,1},\ c_{i,2},\ldots,\ c_{i,K_i})$;
- A pooled detected click rate, r_i, for each group i, where:

$$r_i = \frac{N_i}{d_i}\,. \tag{2.3}$$

Table 2.1: The data available for each group. For illustration purposes only the data for the first 5 groups are shown.

gID	***cs***	***conf***	$\boldsymbol{max_i}$	$\boldsymbol{m_i}$	$\boldsymbol{K_i}$	$\boldsymbol{d_i}$ **(in** μ**s)**	$\boldsymbol{N_i}$	$\boldsymbol{r_i}$	***wisk***	***direc***
1	2	2	740	335.5	6	24.20	2013	83.2	1	0
2	2	1	575	239.7	6	13.67	1438	105.2	1	0
3	3	1	5263	924.0	11	39.40	10164	257.9	1	0
4	2	1	3214	491.0	10	41.95	4916	117.2	0	1
5	5	1	3852	1140.1	9	31.97	10261	320.9	0	1

Since the Whiskey hydrophones are more densely distributed, dives occurring around these may result in groups which are detected by a higher number of hydrophones. This may introduce confounding, hence an indicator (*wisk*) to account the variable "Whiskey" was defined. Additionally, because hydrophones with different directionality may have dissimilar detectability, a binary indicator variable (*direc*) was defined to further investigate that possibility.

Table 2.1 columns represent the potential covariates stated above plus the dependent variable group size (cs), as well as a few indicator variables:

- ***gID*** - the group identification (ID);
- ***cs*** - the cluster (group) size, the dependent or response variable;
- ***conf*** - the confidence level associated with the visual confirmation of group size. This indicator takes the values *conf* = 1,2,3, where 1 = more certain, 2 = more or less certain, and 3 = less certain;
- ***wisk*** - if there is at least one Whiskey hydrophone involved (wisk = 1) or not (wisk = 0);

- ***direc*** - if there is at least one Bidirectional hydrophone involved (direc = 1) or not (direc = 0).

The majority of the groups involved, 43 out of 51, have a confidence level of 1 (*conf = 1*), while 7 groups have a confidence of 2 (*conf = 2*), and only one group has confidence level 3 (*conf = 3*). To evaluate the influence of certainty in group size assignment in the model, another analysis with only groups with a confidence level of 1 will be considered. If the models are not significantly different, and for the sake of using the maximum available data to parametrize a model for predicting group size, all the groups will be used to create the final model.

2.1.2 Density Estimation Dataset

A second dataset (with unknown group sizes) will be used to estimate group sizes based on the model that related group size to acoustic footprint, which will then allow the density estimation over time. It contemplates three different time periods from 2011, covering a total of 113 days. However, 4 out of 113 these days were only partially sampled. Given our objective of producing density estimates per day, it is simpler to consider only the 109 days for which there are 24 hours of recording, and hence these incomplete days were discarded from further analysis.

The considered time periods are: (1) from the 28th of April to the 27th of June; (2) from the 20th of October to the 6th of November; and (3) from the 2nd to the 31st of December, as represented in table 2.2.

Table 2.2: Time periods summary table, discarding the "half-days" and only considering 109 days.

Period	Start date	End data	Total days
1	28/04/2011	27/06/2011	61
2	20/10/2011	06/11/2011	18
3	02/12/2011	31/12/2011	30

2.1.2.1 Raw data

This dataset has a total of 70865 observations, where each line refers to the detections for a given group on a single hydrophone. Hence, each group includes as many rows as hydrophones it was detected on. The same automated procedure that originated the dataset for modelling was also used to obtain this dataset for predictions.

Table 2.3: The data available from the second dataset. For illustration purposes only the data for the first 10 lines are shown.

gID	***edge***	***hyd***	***clickcnt***	***start* (in days)**	***end* (in days)**	***ici* (in secs)**
1	1	93	2058	15091.859392	15091.866419	0.2844
1	1	92	263	15091.859441	15091.866395	0.3416
1	0	90	152	15091.859450	15091.866078	0.2546
1	0	89	85	15091.859635	15091.866173	0.3291
2	1	42	963	15091.867232	15091.887353	0.3575
2	0	43	239	15091.867716	15091.887377	0.3386
3	0	36	2499	15091.867396	15091.895267	0.3767
3	0	37	1500	15091.869670	15091.894790	0.3829
3	1	35	11	15091.872627	15091.873743	0.2540
3	1	35	508	15091.876053	15091.894321	0.3396

Table 2.3 contains the first ten lines from the dataset, where each column corresponds to:

- ***gID*** - the number of the group detected;
- ***edge*** - whether or not the hydrophone that detected the clicks is located at the edge;
- ***clickcnt*** - the number of clicks counted by the corresponding hydrophone;
- ***start*** - time at which the corresponding hydrophone first detected the clicks, in days, where 0 days would correspond to midnight on the 1st of January, 1970 (00h00 UTC, 01/01/1970);
- ***end*** - time at which the corresponding hydrophone ceased detecting clicks, in days, with an identical format as the *start* column;
- ***ici*** - a mean value for the inter-click interval at the corresponding hydrophone.

2.1.2.2 Data Cleaning

The dataset reported above (section 2.1.2.1) was first processed to construct a database with a similar format as that used for modelling, obtaining all the relevant variables required (e.g., the K_i is the number of rows a click was recorded on, the N_i the sum of the click count across those same rows, and so on). Additionally, several procedures to eliminate false positive detections were employed. The following steps were taken sequentially:

1) Hydrophone Duplicates

First, it was necessary to remove multiple records of clicks detected for the same group and same hydrophone, which were created when there were large time gaps between successive click trains in a given hydrophone. To do so, a unique identifier for each row was created, consisting in the group and the hydrophone separated by a dot (e.g. "1.89" corresponds to clicks from the first group detected on hydrophone 89). Records with the same indicator would correspond

to records for the same group and hydrophone, and hence were merged. The values kept were the earliest start and end time, the summation of the click counts, and the minimum inter-click interval. The remaining hydrophone related variables were kept unchanged, since corresponding to records from the same hydrophone, they would have the same value.

2) Click Duration

The next step was to incorporate a new column that refers to the click duration. It is achieved simply by subtracting the *end* and *start* columns.

3) Whiskey Hydrophones

It is also relevant to distinguish between Whiskey and non-Whiskey hydrophones, for which a binary column (0=non-Whiskey, 1=Whiskey) was added. An analysis comparing the potential detection differences between these two types of hydrophones was implemented (see details below).

4) Uni/Bi-Directional Hydrophones

Since the type of beam direction may influence click detection, it makes sense that Uni and Bi-directional hydrophones should be distinguished. A binary column was added (0=Uni, 1=Bi).

5) Data per Group

Since this study addresses the size and density estimation for each *Md* group, there was the need to restructure the data set, such that each record correspond to a single group. To achieve it, the data for each group, comprising as many rows as hydrophones it had been detected on, was condensed into a single row per group with variables at the group level. These included:

- A new column (*nhyd*) was created, with the number of hydrophones (K_i) involved on each group clicks detection;
- A new indicator for the *edge* column was created. If all hydrophones for a group were edge, then the indicator variable edge becomes 1, else it becomes 0. If there is at least a non-Edge hydrophone, it suggests that the corresponding group is most certainly inside the area over which density will be estimated. If $edge = 1$ we assume the record most likely corresponds to a false positive;
- The variable (*shyd*) was also created. If the group was only detected on a single hydrophone, the variable *shyd* was recorded as 1, and 0 otherwise. The former are considered background noise and were removed since it is highly unlikely, if not impossible, for a *Md* individual to pass through the AUTEC range with only a hydrophone detecting the corresponding echolocations. Note therefore a 1 suggests a false positive;
- Since there is the possibility of background noise being detected, hence resulting in false positives, a minimum number of detected clicks per group threshold was set (*thres*). A threshold of 400 clicks was previously tested to be a reasonable amount to consider (Moretti, D., personal communication). Variable *thres* was set to 1 if the total number of

clicks detected for the group was less to 400, and 1 otherwise. Note therefore a 1 suggests a false positive.

6) Additional information

Some additional information was added for each group, such as which period each group belongs to, and the corresponding time and date.

7) Removing False Positives

Lastly, a final column (*est*) was created, which will take the value 1 if the group is to be considered for estimation of density, and 0 otherwise. The variable takes the value 1, meaning the row corresponds to a valid group size and not a false positive, only if none of the 3 false positive indicator variables were true, and 0 otherwise.

Table 2.4 illustrates the first lines of the filtered and transformed data, where each column represents for each group:

- ***gID*** - the number of the group detected;
- ***nhyd*** - the total number of hydrophones the group was detected on;
- ***nclicks*** - the total number of clicks detected by the group;
- ***start*** - time of the first click detected;
- ***end*** - time of the last detected click;
- ***mici*** - the minimum inter-click interval;
- ***edge*** - indicator for whether the group was only detected on edge hydrophones;
- ***shyd*** - indicator for groups detected on a single hydrophone;
- ***thres*** - indicator of whether a minimum number of clicks (400) was detected;
- ***est*** - indicator of whether the group is to be used for density estimation;
- ***period*** - the period (1, 2 or 3) the corresponding group was detected on.
- ***cdur*** - the total duration, in minutes, the hydrophones detected the corresponding group's clicks;
- ***crate*** - a pooled detection click rate for the corresponding group;
- ***date*** - the date each group was first detected, format day/month/year;
- ***jday*** - the julian day of year for the first click detected;

Table 2.4: The data for group size estimation. For illustration purposes only the data for the first 10 lines are shown.

gID	*nhyd*	*nclicks*	*start*	*end*	*mici*	*edge*	*shyd*	*thres*	*est*	*period*	*cdur*	*crate*	*date*	*jday*
1	4	2558	15091.86	15091.87	0.25	0	0	0	TRUE	1	10.12	252.7948	27/04/2011	117
2	2	1202	15091.87	15091.89	0.34	0	0	0	TRUE	1	29.01	41.4357	27/04/2011	117
3	12	14462	15091.87	15091.90	0.25	0	0	0	TRUE	1	40.13	360.3407	27/04/2011	117
5	5	11250	15091.87	15091.91	0.28	0	0	0	TRUE	1	48.64	231.2828	27/04/2011	117
8	8	5430	15091.89	15091.92	0.26	0	0	0	TRUE	1	51.56	105.3188	27/04/2011	117
11	8	6342	15091.92	15091.95	0.23	0	0	0	TRUE	1	39.67	159.8783	27/04/2011	117
14	7	10117	15091.95	15091.97	0.34	0	0	0	TRUE	1	31.09	325.4444	27/04/2011	117
16	12	13149	15091.95	15091.99	0.24	0	0	0	TRUE	1	56.29	233.5776	27/04/2011	117
19	7	6654	15091.97	15092.00	0.29	0	0	0	TRUE	1	37.99	175.1377	27/04/2011	117
20	5	9401	15091.98	15092.00	0.33	0	0	0	TRUE	1	36.80	255.4375	27/04/2011	117

2.2 Some Insights on Exploratory Data Analysis

Before implementing any models, it is useful to thoroughly understand the data. Considering the modelling dataset, an univariate analysis for each explanatory variable was implemented, followed by studying the correlation between them. Some techniques to evaluate correlation between variables are presented next.

Additionally, the data may hold potential differences between the echolocations collected from the several types on hydrophones. Therefore, graphics and histograms were useful to provide additional insights about the data. Each covariate was also studied individually.

2.2.1 Correlation

Examining possible relations within the pool of independent variables is a first step to understand how the variables interact with each other. Analysing the respective graphics may be an important tool to visualize such patterns. Although correlation may take several forms, such as a quadratic pattern, the most common and perceptible correlations happen when a variable increases or decreases linearly or monotonically along with another one. If a variable increases when another ones does, then these two variables are said to be positively correlated. On the other hand, when a variable increases and another decreases, then they are said to be negatively correlated. In the case of a low or no correlation, no discernible linear pattern is present between the two variables.

There are different ways to measure the correlation between variables, even if they differ in nature. Bellow are presented methods, based on correlation coefficients, that take into account the two types of variables the two datasets hold: continuous and dichotomous. All the methods presented are accompanied with significance tests for the correlation coefficients, based on the sample correlation coefficient, considering a level of significance (α) of 0.05, where the null hypothesis refers to no correlation (correlation coefficient = 0).

2.2.1.1 Pearson Product-Moment Correlation Coefficient

According to Cramer (1998), the Pearson product-moment correlation coefficient, or Pearson's ρ, arguably the most widely correlation statistic used, measures the strength of linear dependence between two continuous variables. This coefficient's estimator, R, varies between -1 and 1, where:

- $R = 1$ means a perfect positive correlation between the two variables (they both increase or decrease together);
- $R = -1$ means a perfect negative correlation between the two variables (one increases as the other decreases);
- $R = 0$ means the two variables do not hold a linear dependency.

The closer R is to -1 or 1, the stronger the association. Pearson's ρ is estimated by the formula:

$$R(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\cdot\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S(X,Y)}{\sqrt{S_X^2 S_Y^2}}, \tag{2.4}$$

where X and Y represent two different random variables; and both X_1, X_2, ...,X_n and Y_1, Y_2, ..., Y_n correspond to the sampled populations from X and Y, respectively. Also, $\bar{X}$ and $\bar{Y}$ correspond to the sample means of X and Y, respectively:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i\,, \tag{2.5}$$

where S(X, Y) is the sample covariance between X and Y; and S_X^2 and S_Y^2 are respectively the sample variances of X and Y.

To be able to use this coefficient, both variables have to be measured on either an interval or ratio scale. It is not needed for them to be both measured on the same scale. Nonetheless, outliers may have a great influence on Pearson's correlations, which is why it is useful to compare this coefficient's result with other methods.

2.2.1.2 Spearman's Rank Correlation Coefficient

According to Conover (1999), the Spearman's rank correlation coefficient, or Spearman's r_s, also varies between -1 and 1, and applies to ranks by measuring not a linear, but a monotonic relationship between two continuous or discrete random variables. A monotonic function may be defined as one that is either entirely increasing (for all x and y such as $x \leq$ y, one has *f(x)* $\leq$ *f(y)*), or decreasing (for all x and y such as $x \geq$ y, one has *f(x)* $\geq$ *f(y)*).

Spearman's r_s is defined as the Pearson correlation coefficient between ranked data and its estimator, R_s, may be obtained through the following formula:

$$R_s = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2}\cdot\sqrt{\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2}}, \tag{2.6}$$

where $R(X_i)$ is the rank as compared with the other X values, $i = 1, 2, 3, ..., n$; and $R(Y_i)$ is the rank as compared with the other Y values, $i = 1, 2, ..., n$.

In case of ties, assign to each tied value the average of the ranks that would have been assigned if there had been no ties.

Spearman's rank correlation coefficient is merely what one obtains by replacing the observations by their ranks and then computing Pearson's correlation coefficient on the ranks. Additionally, contrarily to Pearson's ρ, Spearman's r_s is used with ordinal data and is robust to outliers (Altman, 1991).

2.2.1.3 Point-Biserial Correlation Coefficient

According to Sheskin (2011), the Point-biserial correlation coefficient, r_{pb}, is a method to study the correlation between a continuous and a dichotomous variable. It is mathematically

equivalent to the Pearson product-moment correlation, also varying between -1 and 1, and can be obtained using the following formula:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{S_{n\text{-}1}} \sqrt{\frac{n_1 \cdot n_0}{n(n-1)}}, \tag{2.7}$$

where $\bar{X}_1$ and $\bar{X}_0$ denote the sample means on the continuous variable X for the data points where the dichotomous variable Y is either Y=1 (group 1) or Y=0 (group 2), respectively; n_0 represents the number of observations for group 2; n_1 is the number of observations for group 1. Finally, $S_{n\text{-}1}$ corresponds to the sample standard deviation, based on $X_1, X_2, ..., X_n$, i.e.:

$$S_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{X})^2}. \tag{2.8}$$

2.2.1.4 Phi Coefficient

Cramer (1946) defines the Phi coefficient, also known as ϕ, as a measure of association between two binary variables and varies between -1 and 1. It is similar to the Pearson product-moment correlation coefficient in its interpretation. This coefficient is calculated considering the marginal and joint distributions from a 2x2 contingency table, as seen in table 2.5.

Contingency tables are a type of table in a matrix format that presents the frequency distribution of the variables. They hold data assorted simultaneously according to several characteristics.

Table 2.5: 2x2 contingency table for two random binary variables, X and Y.

	$Y = 1$	$Y = 0$	total
$X = 1$	n_{11}	n_{10}	$n_{1.}$
$X = 0$	n_{01}	n_{00}	$n_{0.}$
total	$n_{.1}$	$n_{.1}$	n

In table 2.5, a 2x2 contigency table is presented. The entries in the cells of the table are the frequency counts, denoted by n_{11}, n_{10}, n_{01} and n_{00}, that sum up to n. The marginal totals are represented by $n_{1.}$, $n_{0.}$, $n_{.1}$ and $n_{.0}$.

Two binary variables are considered positively associated if most of the data falls along the main diagonal. On the contrary, they are considered negatively associated if the majority of the data falls off the main diagonal.

The ϕ is defined as:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}. \tag{2.9}$$

2.2.2 Pearson's χ^2 Tests

Pearson's χ^2 tests are commonly used for goodness-of-fit, independence, and homogeneity testing, depending on the type of data one has available and on the sampling design.

Bellow are presented two types of Pearson's χ^2 Tests: one for independence and another for goodness-of-fit.

2.2.2.1 Pearson's χ^2 Independence Test

Greenwood and Nikulin (1996) define Pearson's χ^2 independence test as a method to evaluate if there is a relationship between two categorical variables by evaluating how likely the differences or similarities between them happen by chance. The test is performed under the null hypothesis that the joint distribution of the cell counts in a contingency table is the product of the row and column marginals or, put in other way:

$\boldsymbol{H_0}$: Column classification is independent of row classification

vs.

$\boldsymbol{H_1}$: Column classification is not independent of row classification

Let A and B be two categorical variables with respectively r (rows) and c (columns) categories. When observed over n individuals, a contigency table as illustrated in Table 2.6 can be built.

Table 2.6: Contingency table for two categorical variables, A and B, with r and c categories, respectively.

	B_1	B_2	...	B_c	
A_1	n_{11}	n_{12}	...	$n_{1.}$	$n_{1.}$
A_2	n_{21}	n_{22}	...	$n_{2.}$	$n_{2.}$
...	...	...	...	...	...
A_r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	n

where n_{ij} is the observed value in cell (i,j), with $i = 1, ..., r$ and $j = 1, ..., c$.

The expected frequency, E_{ij}, corresponds to the expected value in cell (i,j). Given the H_0 hypothesis of independence, E_{ij} is calculated as:

$$E_{ij} = n\ p_{i.}\ p_{.j}\,, \tag{2.10}$$

where:

$$p_{i.} = \frac{n_{i.}}{n} = \sum_{j=1}^{c} \frac{n_{ij}}{n}\ ,\ i = 1, ..., r\,; \qquad p_{.j} = \frac{n_{.j}}{n} = \sum_{i=1}^{r} \frac{n_{ij}}{n}\ ,\ j = 1, ..., c\,, \tag{2.11}$$

with $n_{i.}$ referring to the observed frequencies from group i; $\mathrm{p}_{.j}$ denotes the column totals of type j observations ignoring the row attribute; and $n_{.j}$ refers to the observed frequencies from group j. $n_{i.}$ and $n_{.j}$ are also known as the marginal totals.

The statistical test considers the difference between expected and observed values, being defined as the following:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij}\ -\ e_{ij})^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}\,. \tag{2.12}$$

Reject H_0 if $\chi^2_{obs} \geq \chi^2_{1-\alpha;(r-1)(c-1)}$, where $\chi^2_{1-\alpha;(r-1)(c-1)}$ represents the quantile with probability $1-\alpha$ from the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. The closer χ^2_{obs} is to zero, the less significant is the independence between the variables. Note, the convergence to a χ^2 is dependent on the fact that no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater (Yates *et al*, 1999).

2.2.2.2 Pearson's χ^2 Goodness-of-Fit-Test

According to Wayne and Cross (2013), a goodness-of-fit test is suitable to analyse if an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution. It tests if there is evidence to reject H_0, that is, to reject the belief that the data follows a certain distribution. The test takes into account whether or not a sample of observed values of some random variable is compatible with the hypothesis that it is drawn from a population of values which follows a certain probability distribution. Such procedure consists of placing the values into mutually exclusive categories or class intervals and noting the frequency of occurrence of values in each category.

It considers:

$\boldsymbol{H_0}$: The data were drawn/do not deviate from a specified distribution

vs.

$\boldsymbol{H_1}$: The data were not drawn/deviate from a specified distribution

Similarly to Pearson's χ^2 independence test, the goodness of fit χ^2 test, is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i \text{ - } E_i)^2}{E_i} \sim \chi^2_{(k-1-r)}\,, \tag{2.13}$$

where n corresponds to the total number of observations (or the number of cells in the considered table); O_i is the observed frequency for the i^{th} observation; and E_i is the expected frequency for the i^{th} observation.

Reject H_0 if $\chi^2_{obs} > \chi^2_{1-\alpha;(k-1-r)}$, where k is the number of classes of the variable considered, and r is the number of estimated parameters. $\chi^2_{1-\alpha;(k-1-r)}$ represents the quantile with probability $1-\alpha$ from a χ^2 distribution with $(k-1-r)$ degrees of freedom.

2.2.3 Interaction

It is also important to evaluate the interaction between variables. It occurs when the effect of one explanatory variable on the response variable may not be the same at all levels of another explanatory variable. That is, the effect of a explanatory variable on another one is not constant as the effect is not equal for different values the variable takes.

In order to test for interaction, it is common to build a model which considers the corresponding variables and their interaction, and then verifying if the interaction is statistically relevant.

2.2.4 Shapiro-Wilk Test

Shapiro and Wilk (1965) describe the Shapiro-Wilk test as a method to determine whether a sample deviates from a Gaussian distribution. Several statistical tests have a normality assumption, and the Shapiro-Wilk test can be used under that context to evaluate if the assumption is reasonable. Considering a random sample X_1, X_2,..., X_n of size n, with some unknown distribution function, $F(.)$, the following hypothesis are tested:

$\boldsymbol{H_0}$: $F(.)$ is a Gaussian distribution function with unspecified mean and variance

vs.

$\boldsymbol{H_1}$: $F(.)$ is not a Gaussian distribution function

The test statistic, W, is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{[n/2]} -a_i (X_{(n+1-i)} - X_{(i)}) \right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \tag{2.14}$$

where $[x]$ is the largest integer $\leq x$; $X_{(i)}$ is the i^{th} order statistic (where $X_{(1)}$ is the smallest value and $X_{(n)}$ is the largest); $\bar{X}$ corresponds to the sample mean; and a_i are constants generated from the means, variances and covariances of independent and identically distributed random variables of size n sample from the standard normal distribution, and are tabulated in Sarhan and Greenberg (1956). Also, in the numerator, the minus sign in front of a_i makes no difference because of the squaring, but is given because a_i for $2i \leq n$ are negative (Shapiro & Wilk, 1965).

Reject H_0, at the level of significance α, if $W < W_\alpha$, where W_α is the respective critical value. If W is close to 1, the sample behaves like a Normal drawn from a Gaussian distribution.

2.2.5 Multicollinearity

Multicollinearity is a common problem in regression models, happening when two or more independent variables are correlated. Although multicollinearity occurs in most data sets, it becomes an issue when there is a high correlation between the variables, and should therefore be investigated. A high level of correlation means that one variable is linearly related to the others with a considerable accuracy, which may lead to imprecise estimates of the model parameters.

One method to examine whether the independent variables may be correlated is by the *Variance Inflation Factor* (VIF) calculation. A VIF for n explanatory variable is obtained using the R-squared value of the regression (a value which indicates how close the data are to the fitted values) of that variable against all other explanatory variables.

Considering k explanatory variables X, the VIF calculation starts by running an OLS which considers each explanatory variable as a function of all the other predictors. Then, the VIF factor is calculated for each explanatory variable X_i, with $i = 1, ..., k$:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{2.15}$$

where R_i^2 represents the regression R-squared value for the corresponding explanatory variable X_i against all the other predictor variables.

Practical experience points out that if any of the VIF values is higher than 10, then multicollinearity can lead to serious problems (Kutner *et al*, 2004). VIFs can help to identify which regressors are involved in the multicollinearity. Their removal from the analysis should be considered.

2.3 Modelling Approach

Choosing the correct modelling approach for a dataset is a challenge. In this case, the data consists of group size counts, and because counts are always non-negative integers, the Poisson distribution is usually the default option. However, in the presence of overdispersion, i.e., when the observed variance is (considerably) larger than the mean, the Negative Binomial distribution may represent a reasonable alternative (Zuur *et al.*, 2009).

2.3.1 Linear Models

Linear Models (LM) attempt to describe a continuous or categorical dependent variable as a function of one (simple linear model) or more (multiple linear model) continuous or discrete independent variables.

According to Rencher and Schaalje (2008), a linear model has the following form, which hold the systematic and random components:

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k}_{\text{systematic}} + \underbrace{\epsilon}_{\text{random}} \ , \tag{2.16}$$

where Y is the response variable; the regressor variables (also known as predictors) are $x_1, x_2, ..., x_k$; β_0 is a constant which represents the intercept; β_j, $j = 1, 2, \ldots, k$, is the regression coefficient and it corresponds to the rate of change in y for one unit change in the respective j^{th} regressor, assuming the remaining $k-1$ regressors are hold fixed; and ϵ is the error term, which includes everything the model does not take into account by considering the deviations that the observed values y have from the fitted model.

In practice, the betas are not known, and hence must be estimated based on the data. The same happens for the error term, which is estimated via the residual term, denoted by e, as explained further.

LM consist of three components:

1. Systematic component - characterized by the k covariates (and the intercept β_0). The linear predictor, η, is given by:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \, ; \tag{2.17}$$

2. Random component - corresponds to the error term which is assumed to follow a normal distribution, with mean zero and a constant variance σ^2:

$$\epsilon \frown N(0, \sigma^2) \ .$$

As a consequence, the response variable Y (conditional on the regressor variables) follows a normal distribution, with mean μ and constant variance σ^2:

$$Y|x_1, x_2, \ldots, x_k \frown N(\mu, \sigma^2) \ ,$$

with the mean value, $\mu \equiv E(Y|x_1, x_2, ..., x_k)$, depending on the values of the k predictors x_j, $j = 1, 2, \ldots, k$, as one would expect:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \, ; \tag{2.18}$$

3. Link function - characterizes the relationship between the random and the systematic components, and is specified via a link function, $g(\mu)$, with:

$$g(\mu) = \eta \ ,$$

whose objective is to provide a connection between μ and η. Comparing both equations 2.17 and 2.18, it is noticeable that $\eta = \mu$. Thus, in LM the link function is the "identity function" because the mean is modelled directly, as seen bellow:

$$\eta = E(Y|x_1, x_2, \ldots, x_k) = \mu \ .$$

While this might seem a rather convoluted explanation, it is general and sets the scene for other models, where other functions besides the identity function can be considered.

The estimation of the parameters of the model, $\beta_0, \beta_1, ..., \beta_k$, can be done by minimizing the sum of the square of the distances, measured vertically, between the observed values and the model. That is, by the ordinary least squares (OLS) method.

Let each of the k predictor variables, x_1, x_2, ..., x_k, have n observations. Assuming $i = 1, 2, \ldots, n$, x_{ji} represents the i^{th} observation of the j^{th} predictor variable. The observations, y_1, y_2, ..., y_n, constitute realizations of the random sample of size n, Y_1, Y_2, ..., Y_n, from a population Y. Thus, the model 2.16 takes the form of:

$$_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \epsilon_i \, , \quad i = 1, 2, \ldots, n. \tag{2.19}$$

Employing matrix notation simplifies all the math underlying the OLS method; so the model 2.19 can be written in matrix notation such as:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \, , \tag{2.20}$$

with

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \textit{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

where $\boldsymbol{Y}$ is a $n \times 1$ vector of random variables; $\boldsymbol{X}$ is a $n \times (k+1)$ matrix containing the information regarding the observations of the k predictor variables; $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector of regression coefficients; and $\boldsymbol{\epsilon}$ is a n-dimensional vector of the errors.

The least square estimates for the coefficients are given by:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}\,, \tag{2.21}$$

where $\intercal$ represents the transpose of the corresponding matrix. $\hat{\boldsymbol{\beta}}$ will result in a column matrix with $k+1$ entries, where the first entry is the estimate of β_0 and the remaining k are the other slope parameters. Detailed information on this topic can be found, for instance, in Montgomery and Peck (1992).

After obtaining the estimates of the model parameters, the fitted values from the linear regression are computed as follows:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}\,, \tag{2.22}$$

where $\hat{\boldsymbol{y}}$ is the n-dimensional vector of the fitted values.

The error estimate (that is, the residual) for each observation, e_i, is then calculated as follows:

$$e_i = y_i - \hat{y_i}\,, \quad i = 1, 2, ..., n\,, \tag{2.23}$$

or, in matrix form:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}\,, \tag{2.24}$$

where $\mathbf{e}$ is a n-dimensional vector of the residuals.

Also, it is important to refer the projection $n \times n$ matrix, $\mathbf{H}$, or hat matrix, as it is crucial when measuring each observation's influence on the regression model (as seen further in section 2.5.2). This matrix also describes the influence each response value has on each fitted value and is defined as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal\,. \tag{2.25}$$

Thus:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}\,, \qquad \boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\,, \tag{2.26}$$

where $\mathbf{I}$ is the identity matrix of order n.

The hat matrix and its properties play a central role in regression analysis. It is symmetric, idempotent, and rank $(\boldsymbol{H}) = k + 1$, with k being the number of covariates.

Thus far, the assumption of the errors' normality has not been used. This assumption is crucial when constructing statistics for testing hypothesis on the model parameters. Other assumptions for inferential purposes have to be made:

1. Homoscedasticity: $var[\epsilon_i] = \sigma^2, \quad \forall_{i=1,\ldots,n}$;
2. Independence: $cov[\epsilon_i, \epsilon_j] = 0, \quad i \neq j$.

Under these assumptions, it can be proved that the least squares estimator of $\boldsymbol{\beta}$ coincided with the maximum likelihood estimator. Detailed information on hypothesis testing on $\boldsymbol{\beta}$ can be found, for instance, in Montgomery and Peck (1992).

As a final remark, while the model may help predict values for the dependent variable Y, one should not use a regression model to make a prediction for a point that is outside the range of the collected data covariates (i.e. the independent variables). That is called extrapolation and one of statistics' "mortal sins": there is no way to know whether the predicted relationship will hold outside the range of predictor values studied.

2.3.2 Generalized Linear Models

According to McCullagh and Nelder (1989), the term generalized linear models (GLM) refers to a large class of models for a continuous/discrete response variable given continuous and/or categorical predictors. Similar to linear models, the data is still expected to be independently distributed, though they differ on several aspects:

- Errors do not need to be normally distributed, though they still need to be independent;
- GLM allow skewed distributions. Although these models accept a non-normally distributed dependent variable Y, they assume it follows a distribution from the exponential family;
- GLM do not assume a linear relationship between the dependent and independent variables, though they do assume linear relationship between the transformed response in terms of the link function and the explanatory variables;
- The homogeneity of variances, i.e., the residuals follow a common distribution with mean 0 and constant variance σ^2;
- GLM typically use maximum likelihood estimation (MLE) instead of OLS to estimate the parameters; hence relying on large samples properties to obtain precise estimators of the model parameters.

GLM also consist of the same three components of LM:

1. Systematic component - this component is characterized the same way as linear models, where the k covariates combine to create the linear predictor, η:

$$\eta = \boldsymbol{x}\boldsymbol{\beta}\,; \tag{2.27}$$

where $\boldsymbol{x}$ is the $1 \times (k+1)$ vector of the covariates.

2. Random component - similar to linear models, it specifies the distribution of the dependent variable Y. GLM assume a distribution from the exponential family, i.e., Y should have a probability density function (PDF) or a probability mass function (PMF) of the following form:

$$f(y|\theta,\phi) = exp\Big(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big)\,, \tag{2.28}$$

where θ and ϕ are parameters, and $a(\cdot)$, $b(\cdot)$, $c(\cdot,\cdot)$ are real known functions. Any density following the form above is an exponential family density, where θ is called the natural parameter, and ϕ is the dispersion parameter. It is worth to mention that the mean μ of the distribution is related to the natural parameter θ by $\mu = E(Y) = b'(\theta)$.

For the normal distribution (the LM case), $\theta = \mu$ and $\phi = \sigma$.

When considering the Poisson PMF with mean μ:

$$f(y|\mu) = exp\Big\{y\,log(\mu) - \mu - log(y!)\Big\}\,. \tag{2.29}$$

which means that, according to equation 2.28: $a(\phi) = 1$, $\theta = log(\mu)$, $b(\theta) = \mu = e^{\theta}$, $\phi = 1$, and $c(y,\phi) = -log(y!)$.

3. Link function - Contrarily to the LM, the mean is not modelled directly, but through a differentiable transformation resorting to the link function $g(\mu)$, which is now chosen according to the distribution under consideration.

   Considering the Poisson distribution, and according to the equation 2.29, the log link function (canonical link) is used:

$$\theta = log(\mu) \equiv \eta\ ,$$

which implies the expected value, μ, is represented as:

$$log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k\,, \tag{2.30}$$

and since $log(\mu)$ of the response variable is a linear function of the explanatory variables, μ takes the form:

$$\mu = e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k} = e^{\beta_0} \times e^{\beta_1 x_1} \times \ldots \times e^{\beta_k x_k}\,. \tag{2.31}$$

One advantage of suitable chosen link functions is that we can force the predictions to be within a given plausible range, i.e. to model counts as positive numbers or proportions in the (0,1) interval (Hardin & Hilbe, 2007).

As referred before, GLM typically uses the MLE to estimate the unknown parameters. An overview of the MLE method in the GLM framework will be given.

Let each of the k predictor variables, x_1, x_2, ..., x_k, have n observations; with the $1 \times (k+1)$ vector $\boldsymbol{x_i} = [1\ x_{1i}\ x_{2i}\ \ldots\ x_{ki}]$, representing the values of the $k+1$ regressor variables for the i^{th} observation, $i = 1, 2, \ldots, n$. The observations, y_1, y_2, ..., y_n, constitute realizations of the random sample of size n, Y_1, Y_2, ..., Y_n, from a population Y.

MLE is grounded on the *likelihood function*, $L(\theta, \phi; y_1, y_2, \ldots, y_n)$, of the sampled data. It gives the likelihood that the random variables assume a particular value $y_1, y_2, \ldots, y_n$. One wants to know from which density (what values of the unknown parameters) this particular set of values most likely have come from. The likelihood function is, therefore, a function of the unknown model parameters, whose values that maximize the likelihood of the observed sample are the *maximum likelihood estimators*.

According to the probability distribution of the dependent variable Y (equation 2.28), the likelihood function of the n random variables Y_1, Y_2, ..., Y_n, as a function of $\boldsymbol{\beta}$, is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right). \tag{2.32}$$

The common mathematical technique to solve the equation 2.32 involves applying a logarithmic transformation to $L(\boldsymbol{\beta})$, which becomes the *log-likelihood*, $\mathcal{L}(\boldsymbol{\beta})$:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}), \tag{2.33}$$

where:

$$\ell_i(\boldsymbol{\beta}) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi), \tag{2.34}$$

is the contribution of the observed value y_i to the likelihood, $i = 1, 2, \ldots, n$.

Under certain regularity conditions, the maximum likelihood estimators for $\boldsymbol{\beta}$ are given by the following system of equations:

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \quad j = 0, 1, \ldots, k. \tag{2.35}$$

These $k+1$ equations are non-linear in the $k+1$ unknown $\boldsymbol{\beta}$ parameters and, thus, there is no analytical solution for obtaining the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$. Therefore, the parameter estimates are obtained by numerical maximization of the log-likelihood.

There are a number of optimization routines that maximize the likelihood as a function of the parameters, as e.g. Newton-Raphson (N-R). Detailed information on this topic can be found in Hardin and Hilbe (2007).

In the case of GLM bespoke code to maximize the underlying likelihood exist based on the N-R method, via the *glm* function in R. The R function *optim*, implementing the N-R method, is responsible for the implementation of the *glm* function, and can be accessed directly for bespoke problems, but several other options exist (e.g. *nlm* in the *nlme* package).

After maximizing equation 2.32 and obtaining the estimates for the parameters, it is then possible to proceed with hypotheses testing on the parameters, model evaluation, such as residual analysis, goodness-of-fit, etc. More detailed information can be found of Hardin and Hilbe (2007).

GLM were the primary model framework used for this study. Nonetheless, we also investigated the use of Generalized Additive Models for comparison, as described in the following section.

2.3.3 Generalized Additive Models

Generalized additive models (GAM) are an extension of the GLM. GAM can be seen as "non-parametric GLM" because the linear (or some other parametric) form which describes the relation between each covariates and the dependent variable can be replaced by a functional form defined by smoothing techniques.

According to equation 2.27, the linear predictor η specifies that the covariates act in a linear fashion, that is, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$.

Hastie and Tibshirani (1986) introduced a more general form of the linear predictor:

$$\eta = s_0 + \sum_{j=1}^{k} s_j(x_j)\,, \tag{2.36}$$

where $s_j(\cdot)$, with $j = 1, 2, ..., k$, are the unspecified (non-parametric) smooth functions. As seen in the GLM case, the dependent variable Y belongs to the exponential family, which means the predictor η is connected to the dependent variable via a link function.

Instead of estimating single parameters, GAM find a general unspecified function that relates the predicted y values to the predictor values. These unspecified functions are estimated using a *scatterplot smoother*, in an iterative procedure called *local scoring algorithm*. Detailed information about this procedure may be found in Hastie & Tibshirani (1990).

Using a GAM may be a good way to evaluate whether GLM is accurate enough to describe the relationship between a set of covariates and a response variable. It was with that in mind that GAM were considered in this study.

2.3.4 Zero-Truncated Models

According to Zuur *et al.* (2009), if zero counts are not a possibility for the data being modelled, then the underlying PDF may need to be adapted to adjust for the excluded zero counts. Zero-Truncated Models (ZTM) are built for that exact purpose, to model data for which the zero value cannot occur.

ZTM should not to be confounded with Zero-Inflated Models (ZIM), which are more commonly applied in ecological research. According to Zuur & Ieno (2016), ZIM are used when the response variable contains more zeros than expected, a common issue in ecology. ZIM

theory suggests that the excess zeros are generated by a separate process from the count values, allowing the excess zeros to be modelled independently; while ZTM implies the absence of the zero counts and changes the probability of the remaining frequencies.

The objective of zero-truncated models is to model the data excluding the possibility of the response variable to be zero. An example allows a better understanding of what is involved: let Y be the number of *Md* individuals counted in a group. Assuming that Y could be assumed to have a Poisson distribution with mean μ, its probability mass function would be:

$$P(Y = y) = \frac{\mu^y \times e^{-\mu}}{y!}\ , \quad y \in \mathbb{N}. \tag{2.37}$$

$P(Y = 0)$, the probability of observing a zero, is given by:

$$P(Y = 0) = \frac{\mu^0 \times e^{-\mu}}{0!} = e^{-\mu}\ . \tag{2.38}$$

Therefore:

$$P(Y = 0) = e^{-\mu}$$
$$1 - P(Y = 0) = 1 - e^{-\mu}$$

Considering a concrete example, suppose $\mu = 3$, one gets:

$$P(Y = 0) = e^{-3} \approx 0.05 \qquad \text{and} \qquad 1 - P(Y = 0) \approx 0.95$$

This means the probability of observing a positive count would be approximately 0.95, while there is approximately a 0.05 probability of observing a zero. In other words, for every 100 groups 5 would expected to have size zero. As one might suspect, there is no such thing as a group of size zero. The problem aggravates for smaller mean values, where the density condenses more around zero, which would imply a higher amount of zero counts. The solution is attained by modifying the distribution and excluding the possibility of a zero observation.

There is, thus, the need to change the probability mass function in such a way that the probability of $y = 0$ is equal to zero. However, in a trivial context, that would mean that the remaining probabilities would sum up to 0.95. By resorting to ZTM, to obtain a valid PDF, the probability of each outcome larger than 0 is divided by $1 - P(Y = 0)$.

Therefore, there is a need to define the conditional probability of Y as being a Poisson, but strictly positive:

$$P(Y = y|Y > 0) = \frac{P(Y = y, Y > 0)}{P(Y > 0)} = \frac{P(Y = y)}{1 - (P(Y = 0))} = \frac{\frac{\mu^y \times e^{-\mu}}{y!}}{1 - e^{-\mu}} = \frac{\mu^y \times e^{-\mu}}{y!(1 - e^{-\mu})}\ , \quad y \in \mathbb{N}. \tag{2.39}$$

Considering $\mu = 3$, the new probability function is then set as:

$$P(Y = y|Y > 0) = \frac{1}{0.95}\left(e^{-2}\ \frac{3^y}{y!}\right), \quad y \in \mathbb{N}\ . \tag{2.40}$$

Figure 2.1 below illustrates the differences between the probability mass functions (PMF) from a Poisson and a zero-truncated Poisson, both with a mean value of 3.

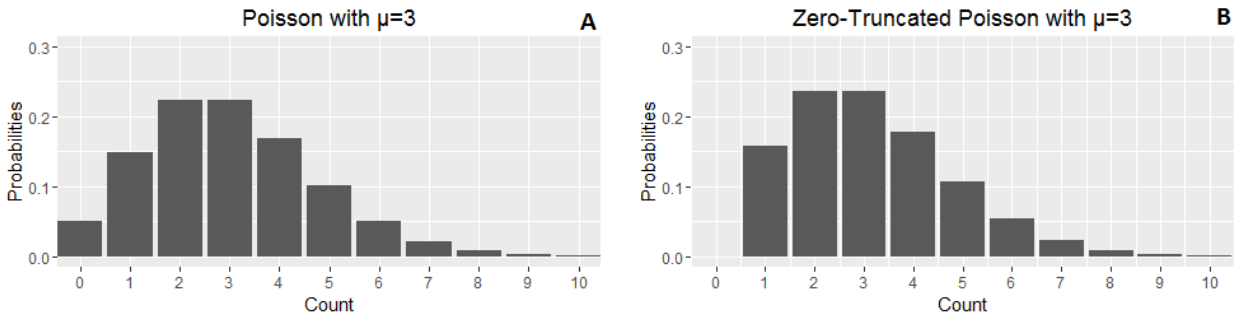


Figure 2.1: **A**: Poisson distribution, $\mu = 3$. **B**: Zero-truncated Poisson, $\mu = 3$, with adjusted probabilities according to Equation 2.40. The vertical lines are slightly higher due to each probability being divided by $1 - P(Y = 0)$. The sum of all probabilities in both A and B is therefore equal to 1, representing a valid distribution.

Summing up, the PMF of the truncated Poisson regression model with k covariates is given by:

$$P(Y = y_i | Y > 0) = \frac{\mu_i^{y_i} \, e^{-\mu_i}}{y_i!(1 - e^{-\mu_i})} \, , \quad y_i \in \mathbb{N}, \quad i = 1, 2, \ldots, n, \tag{2.41}$$

where y_i is the i^{th} observed value of the random variable Y; x_{ji} represents the i^{th} observation from the j^{th} covariate; and:

$$log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}.$$

Note that, while zero truncated models are by far the most common case of truncated models, one could easily extend the framework to deal with any kind of truncated model (e.g., counts must be larger than K_1, or lower than K_2, etc). For the current study, the zero-truncated feature was applied resorting to the *VGAM* package for *R* (Yee, 2015), which incorporates zero-truncation in models for both GLM and GAM, with the commands *vglm* and *vgam*, respectively.

When investigating the best model for further inference, one may consider several model selection criteria. On the next section it will be described how such selection might be implemented, as well as how the best model is chosen.

2.3.5 Modelling with GLM & GAM

Standard GLM and GAM account for zero truncation. However, to compare results, and for the sake of academic curiosity, an *ad hoc* solution which allows the use of non-truncated GLM and GAM was applied: to consider modelling $Y = X$-1, where X is the actual group size. This implies that a new variable (*cs0*) with the transformed response needs to be created, as seen in the table 2.7 below.

Table 2.7: The data available for each group, after adding the *cs0* column. For illustration purposes only the data for the first 5 groups are shown.

gID	***cs***	***conf***	max_i	m_i	K_i	d_i (in μs)	N_i	r_i	***wisk***	***direc***	***cs0***
1	2	2	740	335.5	6	24.20	2013	83.2	1	0	1
2	2	1	575	239.7	6	13.67	1438	105.2	1	0	1
3	3	1	5263	924.0	11	39.40	10164	257.9	1	0	2
4	2	1	3214	491.0	10	41.95	4916	117.2	0	1	1
5	5	1	3852	1140.1	9	31.97	10261	320.9	0	1	4

This means that, while using non-truncated GLM and GAM, the response variable is the *cs0* column instead of *cs*.

2.4 Modelling Strategy: Variable Selection

On any modelling exercise, choosing a suitable model is fundamental. It is desirable for a model to follow the *principle of parsimony*, i.e., to incorporate the minimum possible number of parameters which reasonably explain the response variable. While it is actually debatable which method should be employed to build the most appropriate model, the literature suggests different approaches. Hence, there is still no consensus on the optimal methodology to address this issue. The researcher must always use knowledge of the underlying problem and common sense when evaluating candidate regressors.

Below are presented some of the most commonly used model selection criteria.

2.4.1 Stepwise Regression

In most practical problems, the researcher has a set a candidate regressors which should include all the factors that influence the dependent variable. However, the actual subset of regressors that must be used in the model needs to be determined. Fitting models with different combinations of the regressor variables is usually considered to find the optimal subset(s) of variables. One of the most popular techniques to attain this goal is the stepwise regression methods (see, for instance, Draper & Smith, 1981). Such method examines regression variables subsets by either adding or deleting regressors one at a time. These procedures can be broadly classified into three categories: forward selection; backwards elimination; and stepwise regression, which is a combination of the forward and backward methods. A detailed description of the stepwise regression methods is given in Montgomery and Peck (1992).

2.4.2 Criteria for Evaluating Subset Regression Models

After selecting several subsets of regressor variables, which constitute the different candidate models, it is crucial to decide which subset is the best one. Two criteria for evaluating and comparing subset regression models will be provided.

2.4.2.1 Likelihood Ratio Test

The likelihood ratio test, LRT, is often used to test the parsimony of two models, as long as one of them is a special case of the other (i.e., nested within the other). It is commonly employed when adding or removing variables one by one from a model.

LRT evaluates if a more complex model is required. A broad explanation of LRT is given next. Consider two models: the first contains q regressor coefficients (reduced or restricted model), and the second contains an additional $k+1-q$ regression coefficients (full or unrestricted model). Let the $(k+1)$-dimensional vector of the regression parameters $\boldsymbol{\beta}$ be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_q \\ \hdashline \boldsymbol{\beta}_{k+1-q} \end{bmatrix}, \tag{2.42}$$

where $\boldsymbol{\beta}_q$ is a q-dimensional vector, and $\boldsymbol{\beta}_{k+1-q}$ is a $(k+1-q)$-dimensional vector. The LRT tests the contribution of the $k+1-q$ subset of the regression variables to the model (i.e., $\boldsymbol{\beta}_q \neq \mathbf{0}$). Thus:

$$H_0\colon\ \boldsymbol{\beta}_{k+1-q} = \mathbf{0} \qquad vs \qquad H_1\colon\ \boldsymbol{\beta}_{k+1-q} \neq \mathbf{0}$$

Let $\hat{\boldsymbol{\beta}}_q$ and $\hat{\boldsymbol{\beta}}$ represent the maximum likelihood estimators for the two models; and L_q, L denote the values of the likelihood functions for the two models evaluated at $\hat{\boldsymbol{\beta}}_q$ and $\hat{\boldsymbol{\beta}}$, respectively.

Under the null hypothesis, the LRT statistic, denoted by D, is given by:

$$D = -2ln\left(\frac{L_q}{L}\right) \sim \chi^2_{(k+1-q)}\,. \tag{2.43}$$

Reject H_0 when $D > \chi^2_{(1-\alpha)(k+1-q)}$, which represents the quantile with probability $(1-\alpha)$ from a χ^2 distribution with $(k+1-q)$ degrees of freedom, where α is the significance level.

When the null hypothesis is not rejected, that means the extra regression variables do not increase the fit enough to justify their inclusion in the model. Adhering to the parsimony principle, the model considering less variables, the reduced model, is a better representation of the data.

2.4.2.2 Akaike's Information Criterion

The Akaike's information criterion, AIC is a model selection criterion and attempts to choose from a group of models the one which appears to be the most accurate model to describe the response variable. The AIC is calculated as:

$$AIC = -2\,log\ L(\hat{\boldsymbol{\beta}}) + 2k\,; \tag{2.44}$$

where $\boldsymbol{\beta}$ is the vector of the regression parameters; $L(\hat{\boldsymbol{\beta}})$ represents the likelihood function evaluated at the maximum likelihood estimate of $\boldsymbol{\beta}$; and k is the number of regression variables.

The more parsimonious candidate model is the one with the lowest AIC. Since this function is multiplied by -2, the model with the lowest AIC is the one with the highest likelihood

function value. An adjustment is made to the AIC equation by adding the number of estimated parameters ($2k$) to this measure. The model is penalized by the number of parameters added: more parameters add to a higher AIC value (Fabozzi *et al.*, 2014).

However, there may be times the AIC values from different models are extremely similar, only differing by a single unit, or even less. How should one proceed when confronted with such issue? According to Burnham and Anderson (2002), models within 1 or 2 units of the best model (the one with the lowest AIC value) have substantial support from the data. However, there is currently some controversy about these specific values (Burnham & Anderson, 2002; Fabozzi *et al.*, 2014).

For this current study, these perspectives were taken into consideration to decide between models, also including an analysis to each model's quality, as described on the next section.

2.5 Residual Analysis and Influential Observations

When conducting statistical analysis it is important to evaluate how well the model fits the data and if the data meet the assumptions of the model; to detect outliers; and to detect influential observations (i.e., observations with a high influence on the fitted model). In section 2.5.1, a summary on residual analysis is presented. Section 2.5.2 provides a few measures on influential observations.

2.5.1 Residuals

It is an usual practice to analyse the residuals from the candidate models, as this provides a measure of goodness-of-fit, how adequate the model actually is for the data at hand. In the LM context, each residual, e_i, corresponds to the discrepancy between the observed value, y_i, and the fitted value, $\hat{y}_i$, with $i = 1, ..., n$, as seen before in equation 2.23. Residuals should be "well-behaved", by showing no distinguishable pattern and a constant variance.

Although, when considering a GLM framework the residuals definition is not unique and, thus, their interpretation is less clear and graphical patterns may vary quite differently for different models. Several residuals definitions have been proposed for the GLM models, and in particular for the Poisson regression. In this section some of those definitions will be presented. However, it is important to emphasise there appears to be no one single residual definition that be used in all contexts. For Poisson regression models, there is no one residual that has zero mean, constant variance, and symmetric distribution. This leads to several different residuals according to which of these properties is felt to be the most desirable (Cameron & Trivedi, 1998).

For academic purposes, a residual analysis still took place, along with other model quality measures as described next.

Raw residuals

From the LM context, the raw residuals are the "natural residuals":

$$e_i = y_i - \hat{\mu_i}\,, \qquad i = 1, 2, \ldots, n\,, \tag{2.45}$$

where the fitted mean $\hat{\mu_i}$ is the conditional mean $\mu_i = E(Y_i|\boldsymbol{x}_i)$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

For count regression models, Cameron and Trivedi (1998) showed that the raw residuals are heteroskedastic and asymmetric.

Pearson Residuals

According to Cameron and Trivedi (1998), the obvious correction for heteroskedasticity is resorting to the Pearson Residuals:

$$p_i = \frac{y_i - \hat{\mu_i}}{\sqrt{\hat{\omega_i}}}, \qquad i = 1, 2, \ldots, n\,, \tag{2.46}$$

where $\hat{\omega_i}$ is the estimated variance ω_i of y_i.

For large samples, these residuals have zero mean, and are homoskedastic (with unit variance), but are asymmetrically distributed.

For the Poisson regression, one gets:

$$p_i = \frac{y_i - \hat{\mu_i}}{\sqrt{\hat{\mu_i}}}, \qquad i = 1, 2, \ldots, n\,. \tag{2.47}$$

Once the residuals are obtained, one should use them to extract fruitful information. For instance, residuals should be plotted against the predicted values of the dependent variable; and against the regressors under study, to see whether regressors should enter through a different functional form than that specified (Cameron and Trivedi, 1998).

2.5.2 Influential Observations

When looking at the residuals, one often finds observations that may influence the model in several ways. For instance, if an observation has a response value which holds a large difference from the predicted value, then that observation may be described as an outlier, i.e., the observation has an extreme or a notably different y value than the rest. Regression outliers usually have large residuals but do not necessarily affect the regression slope coefficient.

On the other hand, if an observation stands out from one or more predictor values, then it is said to have a high leverage. In other words, leverage measures how unusual that point is when comparing with all the other observations. High leverage does not necessarily mean the observation will influence the regression coefficients, i.e., it may have an extreme value when compared to the other points, but following at the same time the prediction tendency (Chatterjee and Hadi, 1986).

In LM, the matrix $\boldsymbol{H}$ is one of the most common measures of leverage. According to equation 2.26, $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. Consider h_{ii} the i^{th} diagonal element of the projection matrix $\mathbf{H}$, $i = 1, 2, \ldots, n$. If h_{ii} is large, then the matrix $\boldsymbol{X}$, which determines $\boldsymbol{H}$, is such that y_i has a large influence on its own prediction (Cameron and Trivedi, 1998).

In the GLM case, Hardin and Hilbe (2007) proved that the $\boldsymbol{H}$ matrix is given by:

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{W}^{1/2}\,, \tag{2.48}$$

where $\boldsymbol{W} = \text{diag}\left\{\frac{1}{V(\mu)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)^2\right\}$, with $V(\mu) = \frac{\partial\mu}{\partial\theta}$. As in LM, the $n \times n$ matrix $\boldsymbol{H}$ is idempotent with trace equal to its rank $k+1$, the number of regressor variables. Therefore, the average value of h_{ii} is $\frac{(k+1)}{n}$, and the values of h_{ii} in excess of $\frac{2(k+1)}{n}$ are viewed as having high leverage (Cameron & Trivedi, 1998; Hoaglin & Welsch, 1978). This will be the criterion used for the present work.

Studentized Pearson Residuals

According to Hardin and Hilbe (2007), the studentized Pearson residuals, p_i^*, are then given by:

$$p_i^* = \frac{p_i}{\sqrt{1-h_{ii}}}, \qquad i = 1, 2, \ldots, n\,. \tag{2.49}$$

Hat values are the most common measure of leverage. They are calculated based on the fitted values from the regression and are defined as:

$$h_{ii} = [\mathbf{H}]_{ii} \ ,$$

where h_{ii} is the i^{th} observation leverage score, which is found in the i^{th} diagonal element of the projection matrix $\mathbf{H}$. Hat values describe the influence each response value has on the fitted value for that same observation.

Although literature considers different cut-off criteria, Hoaglin & Welsch (1978) suggest h_{ii} is a high leverage point when $h_{ii} > \frac{2p}{n}$, where p is the summation of the hat values, and n is the sample size. This will be the criterion used for the present work. In simple regression, hat values measure the distance of each points from the expected value. In multiple regression, the distance from the centroid point.

Cook's Distance

Another way of measuring an observation's influence is by calculating its Cook's distance, D_i, which measures how much the regression estimated coefficients change when the i^{th} observation is removed from the estimation procedure. In the GLM context, Hardin and Hilbe (2007) approximate Cook's distance with:

$$D_i = \left(\hat{\boldsymbol{\beta}}^*_{(i)} - \hat{\boldsymbol{\beta}}\right)^\intercal \mathbf{I} \left(\hat{\boldsymbol{\beta}}^*_{(i)} - \hat{\boldsymbol{\beta}}\right) , \tag{2.50}$$

where $\mathbf{I}$ is the Fisher information matrix; $\hat{\boldsymbol{\beta}}^*_{(i)}$ is the one-step approximation to the jackknife-estimated coefficient vector.

These measures can be used to detect observations with undue influence. An observation i is considered to have high influence when $D_i > \frac{4}{n-k}$ (Hair *et al*, 1998; Hardin & Hilbe, 2007). Observations with measures greater than $4/n$ should be investigated (Hardin & Hilbe, 2007).

Even if no observations exceed these thresholds, additional attention should be dictated if a small set of observations has substantially higher values than the remaining observations (Hair *et al*, 1998).

This study contemplated these issues, so to mitigate any extra weight from the influential points, different models were built taking into account a new dataset were these observations were removed. The different models were then compared, and if they held no major difference then no observations were removed when modelling.

2.6 Group Size and Density Estimation

Before proceeding with the estimation, an exploratory analysis was performed on the second dataset. Building on the best model chosen using the tools and methods described above, the group size for each dive observation in the density estimation dataset was then predicted.

After estimating the group sizes we can finally estimate *Md* density. Following Moretti *et al.* (2010), the estimator of animal density is given by:

$$\hat{D} = \frac{n\ \hat{s}}{\hat{r}\ TA}\,; \tag{2.51}$$

where:

- $\boldsymbol{n}$ - total number of dives/groups;
- $\hat{s}$ - estimated average group size (common to all groups);
- $\hat{r}$ - estimated average number of dives per hour (a value of 0.36 measured by Moretti *et al*, 2010);
- $\boldsymbol{T}$ - considered amount of time;
- $\boldsymbol{A}$ - considered area (1291 km^2 in the current study).

If the area (A) is removed from the equation, the estimator of abundance ($\hat{N}$) is obtained.

$$\hat{N} = \frac{n\ \hat{s}}{\hat{r}\,T}\,; \tag{2.52}$$

The abundance corresponds to the total amount of animals detected for the time period considered, whereas the density is simply the abundance taking into account a certain area.

While Moretti *et al.* (2010) use an average estimated value ($\hat{s}$) common to all groups, this work focuses on the previously obtained model to estimate the number of individuals for each detected group. Instead of resorting to the total number of dives and multiplying that value for an estimated average group size based on literature, this study suggests a more precise approach for *Md* density estimation. Therefore, and estimator of density ($\hat{D}$) is now obtained using a different equation:

$$\hat{D} = \frac{\sum_{i=1}^{n} \hat{s}_i}{\hat{r}\,TA}\,; \tag{2.53}$$

where $\hat{s}_{\boldsymbol{i}}$ corresponds to the estimated group size for group i, and n represents the number of groups for the considered time period. For this study, the density was estimated for each day, considering the amount of *Md* individuals in 1000 km^2. In fact, to make the analogy with equation 2.52 , we can represent the density per day ($\hat{D}_d$) as a function of the mean group size per day:

$$\hat{D}_d = \frac{\sum_{i=1}^{n_d} \hat{s}_i}{\hat{r}\, TA} = \frac{n_d\, \hat{\bar{s}}_i}{\hat{r}\, TA}\,; \tag{2.54}$$

where n_d represents the number of groups on day d; $\hat{s}_i$ is the estimated mean group size on day d; and $\hat{\bar{s}}_i$ is the estimated group size mean.

A new table was created pooling information for each day (table 2.8), including:

- ***groups*** - the number of groups detected on that day;
- ***mcs*** - a mean cluster (group) size value for the corresponding day;
- ***stime*** - time (in hours) when the first group was detected for the corresponding day;
- ***etime*** - time (in hours) when the last group ceased being detected for the corresponding day;
- ***ttime*** - the time frame (in hours) per day the click detection occurred, i.e., the time period over which the measurement was made;
- ***abundance*** - the abundance of individuals detected per hour on the corresponding day;
- ***density*** - the density of individuals detected per hour on the corresponding day, per 1000 km^2 of the total considered area.

Table 2.8: The first ten lines from data regarding each day, featuring the estimated abundance and density.

day	*groups*	*nhyd*	*nclicks*	*ici*	*period*	*crate*	*mcs*	*stime*	*etime*	*ttime*	*abundance*	*density*
117	10	70	80565	0.2315	1	214.0648	3.3889	20.6167	23.4000	2.7833	33.8218	26.1981
118	58	339	383357	0.2300	1	165.2208	3.1689	1.0000	23.4833	22.4833	22.7083	17.5897
119	28	136	128323	0.2310	1	106.4493	2.8435	0.9833	23.3167	22.3333	9.9028	7.6707
120	58	325	322389	0.2313	1	148.3170	3.0232	0.9833	23.2167	22.2333	21.9073	16.9693
121	67	379	352552	0.2301	1	147.9339	2.9175	0.8667	23.4000	22.5333	24.0969	18.6653
122	57	322	292268	0.2303	1	144.8485	2.7951	1.01667	22.8500	21.8333	20.2697	15.7008
123	74	413	376923	0.2300	1	129.2886	2.8160	1.1167	23.9500	22.8333	25.3509	19.6366
124	92	492	496524	0.2318	1	120.0953	2.8769	0.2333	23.6833	23.4500	31.3529	24.2858
125	82	502	476841	0.2307	1	132.0471	2.8462	1.0333	23.4333	22.4000	28.9419	22.4182
126	42	197	207308	0.2307	1	117.4795	3.0406	1.3000	23.0667	21.7667	16.2972	12.6237

2.6.1 Bootstrap

One main goal of inferential statistics is to determine the value of a population parameter. According to Manly (2006), bootstrap is a form of random statistical sampling which evaluates the precision of the sample estimates by estimating properties of those same estimators. When computing a statistic from a single dataset only that one statistic is known. There is no information about the variability of that statistic. That is why bootstrap is helpful, since it creates a large number of possible datasets, and computes the statistic on each of these datasets; thus providing a distribution of the statistic. That is the strategy of bootstrap: to create data that "might have been seen" (Chernick & LaBudde, 2011).

The basic idea behind this method starts with a sample with size n, which only allows one estimate of the parameters say, the mean, or the variance. Then, this same sample is randomly re-sampled with replacement to build a new sample, also with size n. Here is the trick: since the re-sample occurs randomly and with replacement, several elements will most likely be repeated in the new sample while some will be missed altogether and, as n increases, the probability that the new sample will look exactly like the original one will tend to zero. This process is then repeated a great number of times, and for each of these new samples the parameter in study will be computed. The variance over these bootstrap pseudo values for the statistic of interest (say, the mean) will be an approximation of the variance of the estimator for that mean.

Considering the present study, the final task is to propagate the variance in the model of group size thorough the estimates of variance of density per day. This is straightforward to do within a bootstrap context. Therefore, the modelling dataset will be re-sampled 999 times. For each re-sample, the model selected for inference will be refit. This will therefore lead to new parameter estimates, and hence, corresponding different predictions for each of the groups sizes one needs to predict. At each iteration the density per day will be calculated. Therefore, in the end, there will be K estimates for each day's density, and getting variance or confidence intervals using the percentile method (Manly, 2006, p. 46-51) from these allows one to obtain precision measures which incorporate the model uncertainty in the final inferences.

2.6.1.1 Parametric Bootstrap

The bootstrap technique mentioned before used the *empirical bootstrap*, which draws bootstrap samples by resampling the data, making no assumptions about the underlying distribution. The difference between the empirical and the parametric bootstraps is the source of the bootstrap sample. The parametric type produces the bootstrap sample from a parametrized distribution by fitting a parametric model to the data, often by MLE, and samples of random numbers are drawn from this fitted model. Confidence intervals for the parameter may then be built.

The bounds for a $100(1-2\alpha)\%$ approximate standard normal confidence interval are given by:

$$\hat{\theta}_L = \hat{\theta} - z_\alpha \cdot \hat{se}(\hat{\theta})\,, \qquad \hat{\theta}_U = \hat{\theta} + z_\alpha \cdot \hat{se}(\hat{\theta})\,; \tag{2.55}$$

where $\hat{\theta}$ is the estimation of the parameter of interest θ; $\hat{\theta}_L$ and $\hat{\theta}_U$ are respectively the lower and upper bounds of the confidence interval of θ; and $z_\alpha = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)^{th}$ quantile of the standard normal distribution.

However, the standard normal confidence interval has the drawback of always being symmetric around the estimated parameter.

Efron and Tibshirani (1993) then suggest replacing the confidence interval with the percentile interval. It is based on the empirical percentiles of the bootstrap replicates. Given B bootstrap samples, the replicates, $\hat{\theta}^*$, are ordered from the smallest to the largest, where the bounds for the confidence interval are chosen from the $B\alpha^{th}$ and the $B(\alpha-1)^{th}$ replicates. I.e., the bounds for a $100(1-2\alpha)\%$ percentile interval are given by:

$$\hat{\theta}_L = \hat{\theta}^*_{(B\alpha)}\,, \qquad\qquad \hat{\theta}_U = \hat{\theta}^*_{(B(1-\alpha))}\,. \tag{2.56}$$

A parametric bootstrap was employed in the present study, since Moretti *et al* (2010) calculated a weighted mean dive rate ($\hat{r}$) of 0.36 dives/hour, with a weighted standard error of 0.04. In order to include the dive rate variance in each bootstrap, a vector with 999 values was created (one for every single bootstrap), from a normal distribution with a mean of 0.36 and a standard error of 0.04.

Chapter 3

Results

The code required to reproduce the results is provided in Appendix A.

3.1 The Modelling Dataset

3.1.1 Exploratory Analysis

An exploratory analysis for the modelling dataset took place before beginning the model selection.

Figure 3.1 reveals the total number of clicks detected for each hydrophone. The salmon and blue bars represent uni and bi directional hydrophones, respectively. It is also important to highlight that the hydrophones numbered 1-14 are Whiskey.

Figure 3.2 illustrates the number of clicks detected for each group. One may see that the groups 13 and 47 stand out from the rest.

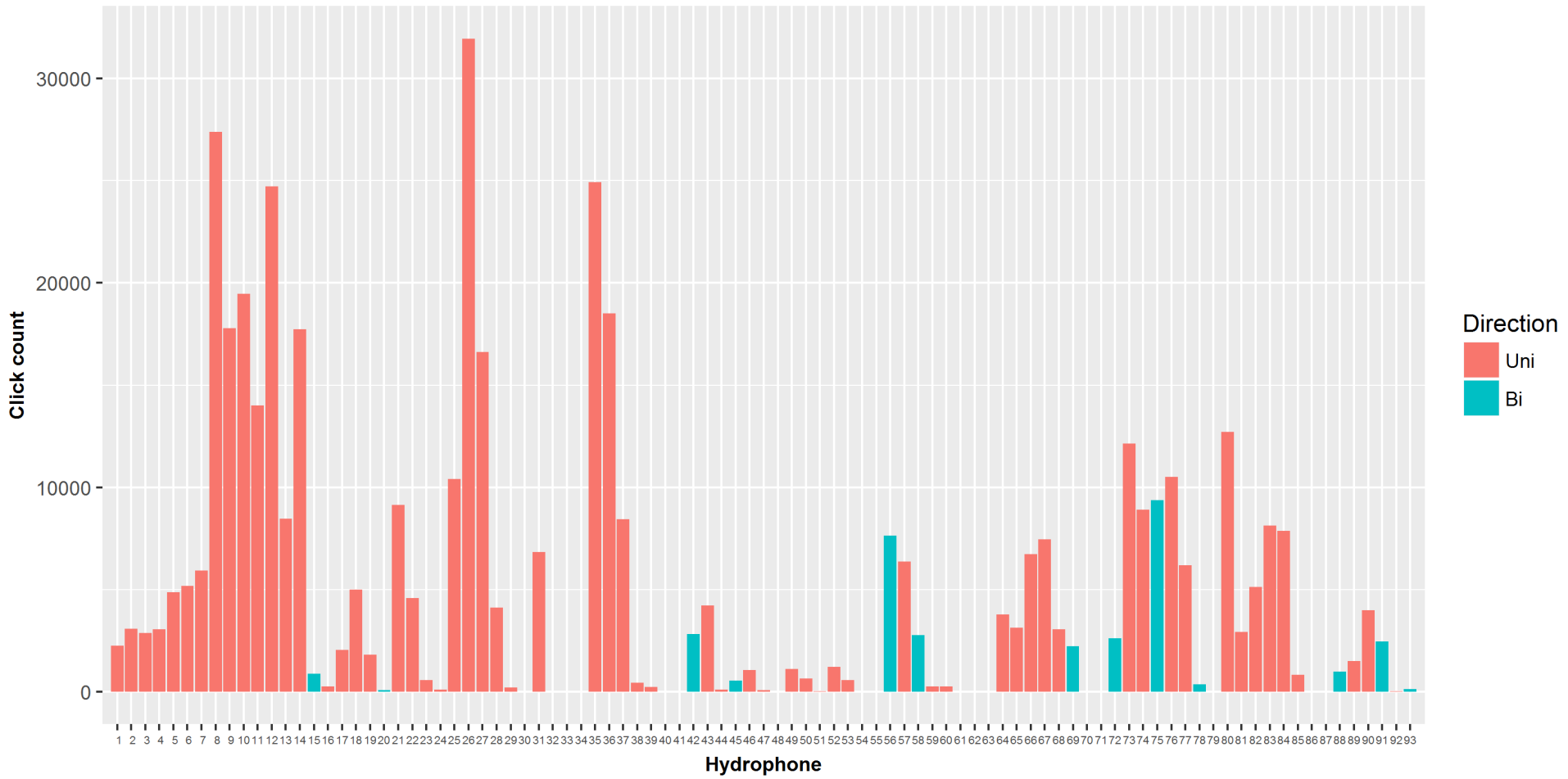


Figure 3.1: Click counts for all 93 hydrophones. Uni and Bi hydrophones are distinguished with different colours (salmon and blue, respectively).

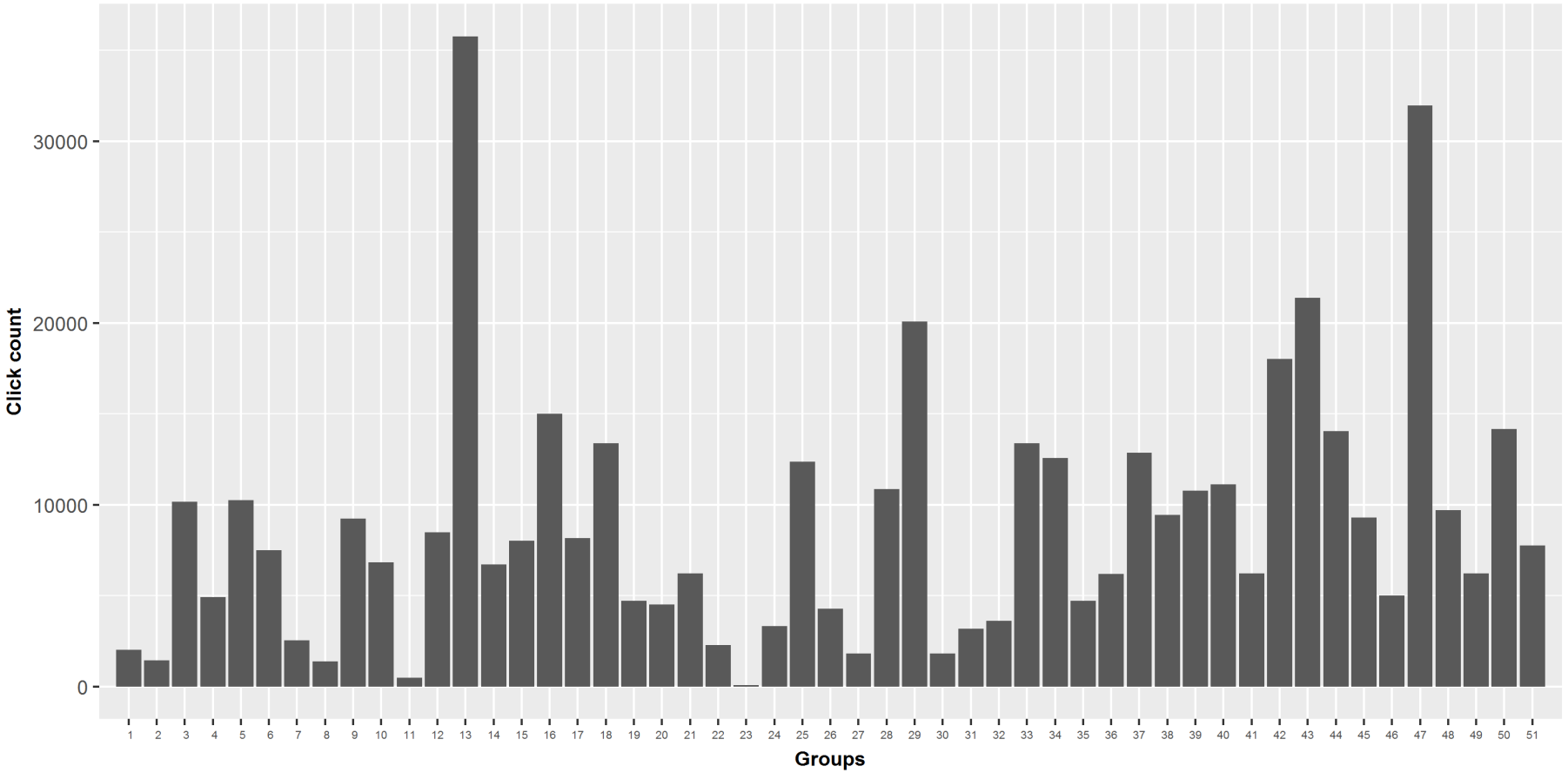


Figure 3.2: Total number of clicks detected for each one of the 51 groups.

In any regression framework it is recommended to understand the distribution of the response variable prior to implementation of plausible candidate models. In our setting, that corresponds to investigate the distribution of observed group sizes for the groups for which group size is assumed to be known.

Figure 3.1 illustrates the counts from the response variable.

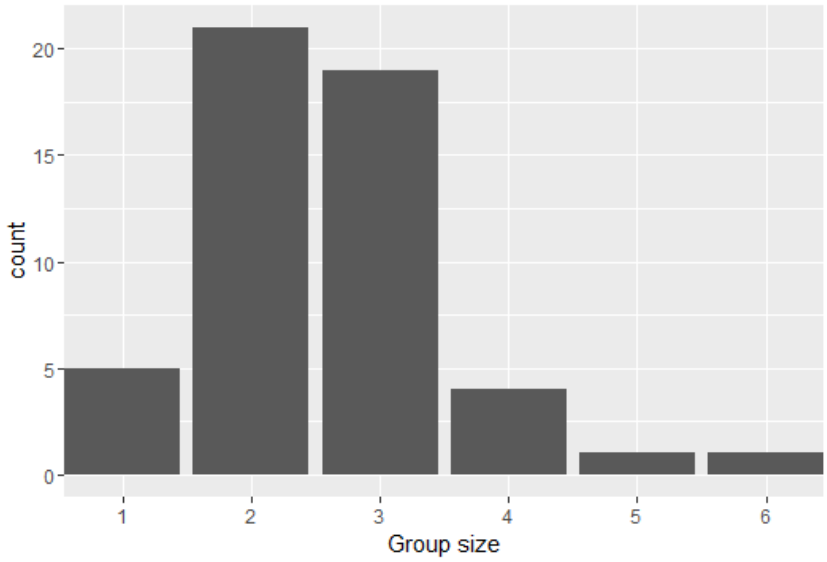


Figure 3.3: Group size count distribution for the modelling dataset, ranging from 1 to 6 individuals per group, with a total of 51 groups.

Ignoring the lack of zeros, the group size distribution appears to resemble a Poisson distribution. Although, before testing that hypothesis, it is important to check for overdispersion. The sample's group size mean is approximately 2.57, while the sample's group size variance is about 0.97. Since $0.97 < 2.57$, it looks like there is no overdispersion, but probably underdispersion instead. Underdispersion occurs when the variance is smaller than the mean, and it is actually a rare case. In many cases it does not constitute an issue, a Poisson may continue to reasonably fit the response variable distribution (Frome, 1982).

To verify the validity of a Poisson distribution, a Pearson's χ^2 goodness of fit test was employed by comparing the group size distribution with a Poisson distribution with a mean value equal to the sample's group size mean. However, since the standard Poisson considers the value zero, an adjustment for this test was made to incorporate the zero in the data group size by removing a single unit to each y value, i.e., by considering the response variable to be $cs0$ instead of cs, similarly as what was described on the previous chapter.

The hypothesis to consider are:

$$\boldsymbol{H_0}\text{: The sample's group size follows a Poisson distribution}$$
$$\textit{vs.}$$
$$\boldsymbol{H_1}\text{: The sample's group size does not follow a Poisson distribution}$$

The following results were then obtained:

X-squared = 9.3333
p-value = 0.1557

However, a warning message appeared: "Chi-squared approximation may be incorrect". A quick fix relies in simulating the p-value based on replicates for the sample size with $n = 51$ (in this case 10000 replicates). The results were:

X-squared = 9.3333
p-value = 0.1540

The null hypothesis, $\boldsymbol{H_0}$, is not rejected for the usualy considered significance values (0.01, 0.05 and 0.1). This means the analysis may proceed considering the Poisson distribution, since it appears to fit well to the data.

3.1.1.1 Univariate Analysis

In this section each explanatory variable's relation with the response variable will be analysed. The plots corresponding to this relation may be seen on figures 3.4 and 3.5:

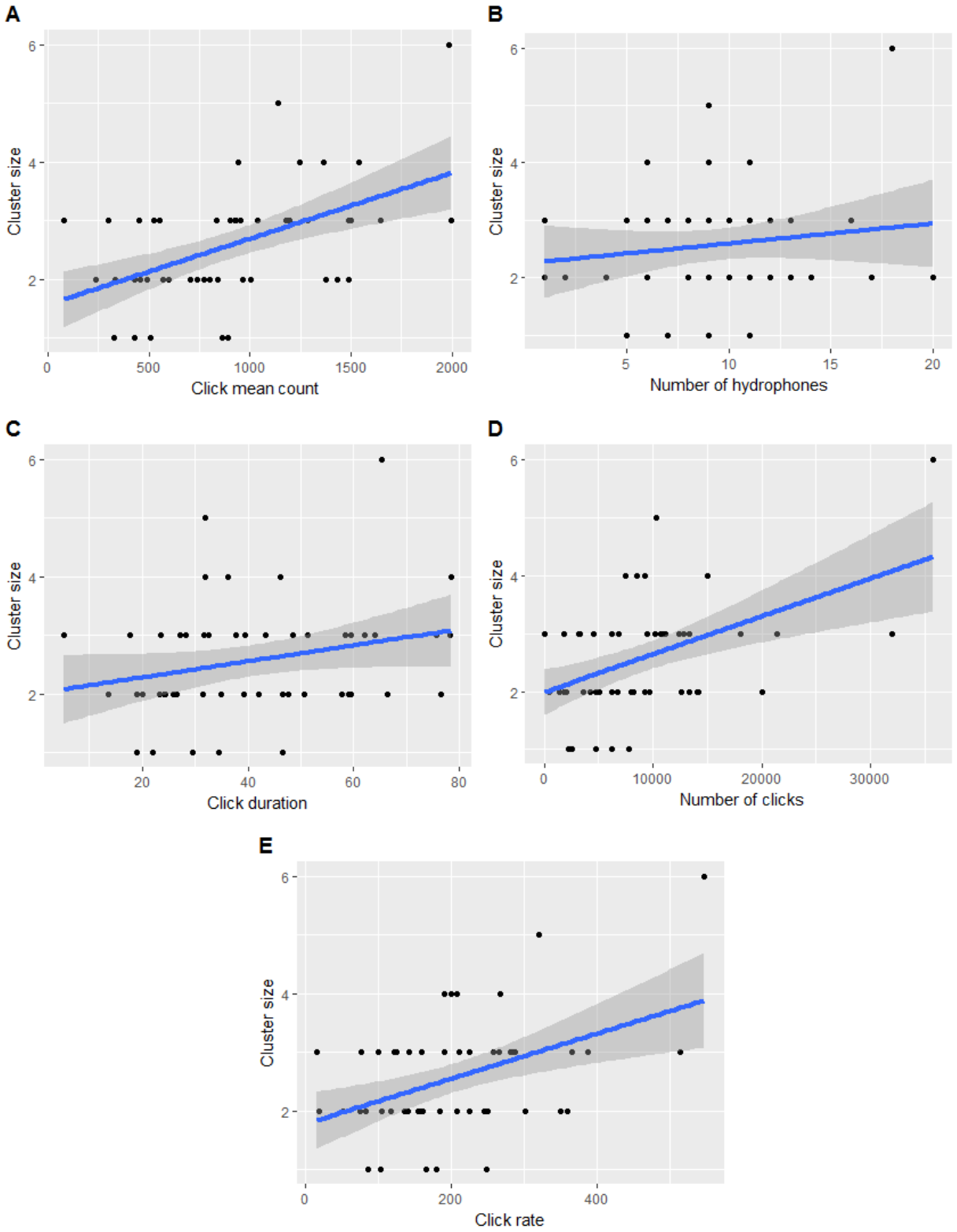


Figure 3.4: Univariate analysis for each continuous explanatory variable (x-axis) against the response variable, cluster size (y-axis). Each black dot corresponds to an observation and the blue line matches the regression line, where the grey area is the 95% confidence level interval for the predictions. **A**: click mean count, **B**: number of hydrophones, **C**: click duration, **D**: number of clicks, **E**: click rate.

The group size appears to increment as each continuous explanatory variable increases. However, only the variables "click mean count", "number of clicks", and "click rate" appear to be statistically significant with a *p-value* < 0.05, when building univariate regression models.

Figure 3.5: Univariate analysis for each binary explanatory variable (x-axis) against the response variable, cluster size (y-axis). Each violin plot considers an orange area where its width is proportional to the number of observations, and a black dot that corresponds to the observations' median. **F**: whiskey/non-whiskey, **G**: uni-directional/bi-directional.

Looking at the binary variables, it appears that group size decreases with Whiskey hydrophones. It seems to indicate that smaller group sizes were more commonly detected on the these hydrophones. On the contrary, groups size shows the opposite behaviour with the variable "direction". Both variables are not statistically significant on their univariate analysis (p-*value* > 0.05).

3.1.2 Correlation

Correlation, whether causal or not, may indicate a predictive relationship that can be beneficial, since it may be possible to predict a variable from another one.

Figure 3.6 illustrates each non-binary variable (**meancount, nhyd, cdur, nclicks, crate**) behaviour against each other.

Figures 3.7 and 3.8 illustrate the correlation between the non-binary variables via Pearson's ρ and Spearman's r_s, respectively.

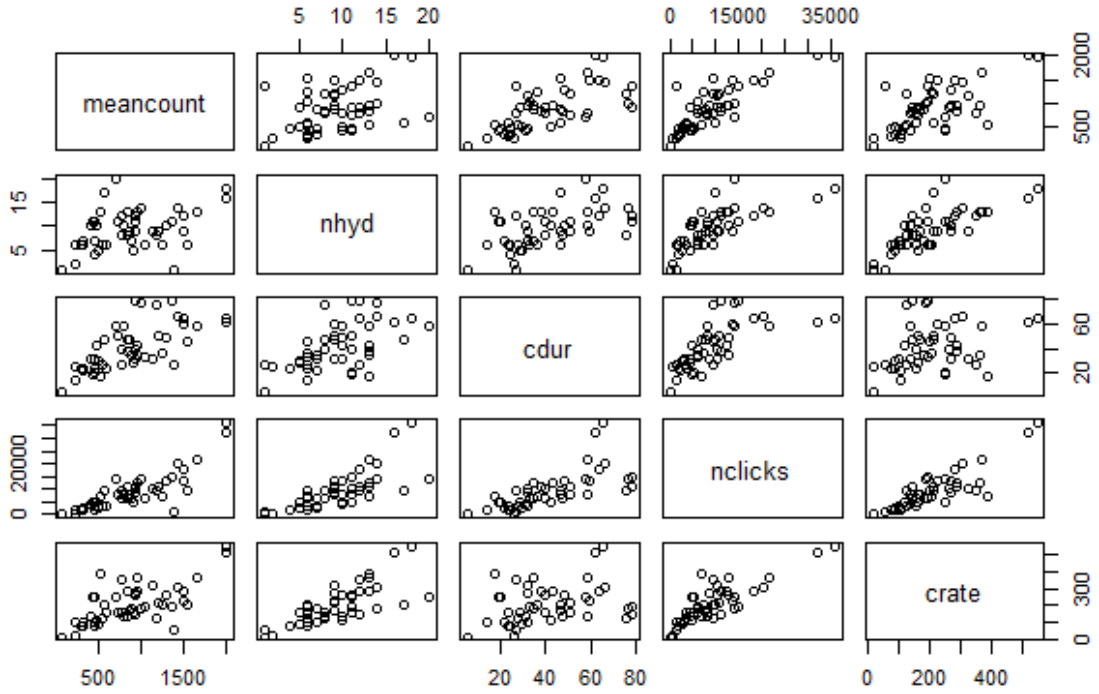


Figure 3.6: Behaviour of each non-binary variable against each other (mean count, number of hydrophones, click duration, number of clicks, and click rate, respectively).

Figure 3.7: Correlation plot featuring Pearson's ρ value for each non-binary variable duo (mean count, number of hydrophones, click duration, number of clicks, and click rate).

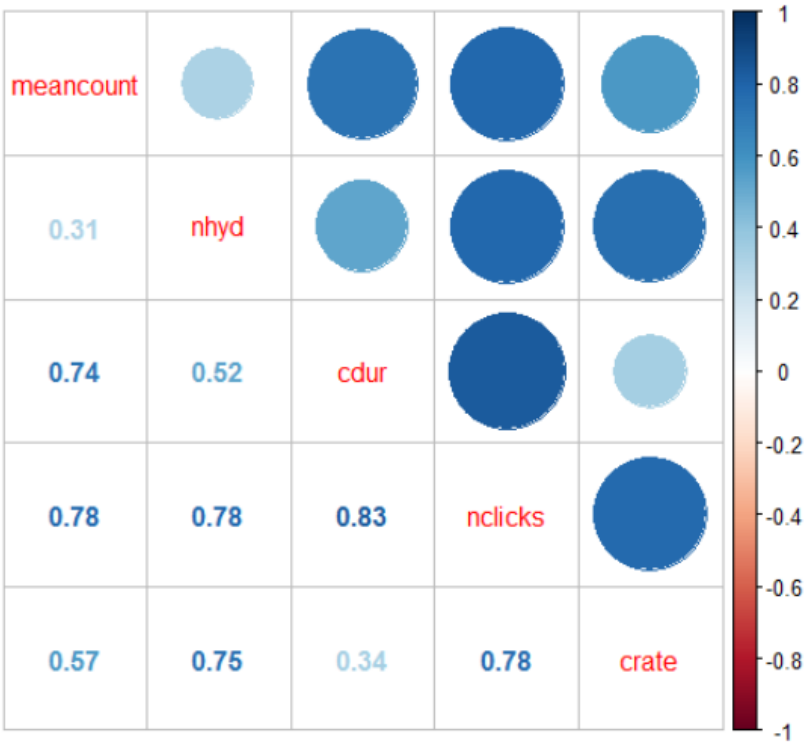


Figure 3.8: Correlation plot featuring Spearman's r_s value for each non-binary variable duo (mean count, number of hydrophones, click duration, number of clicks, and click rate).

Figure 3.9 exhibits the Point-Biserial coefficients between the non-binary and binary (**direction, wisk**) variables.

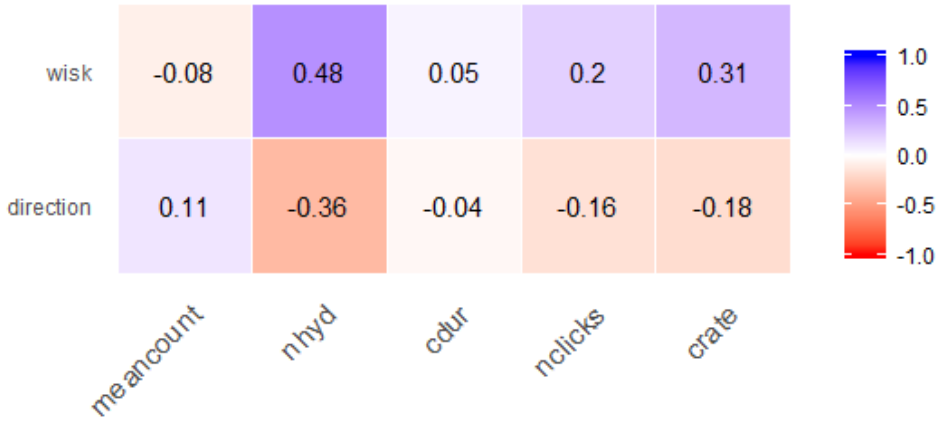


Figure 3.9: Correlation plot featuring the Point biserial correlation coefficient between the non-binary (mean count, number of hydrophones, click duration, number of clicks, and click rate) and the binary (direction and whiskey) variables.

Finally, when using the Phi coefficient measure between both binary variables, one gets $\phi = -0.68$.

3.1.3 Model Building

After getting an insight into the modelling dataset, it is time to build candidate models that would explain the group size. Standard Poisson GLM and GAM were the first models employed, followed by the zero-truncated approach.

3.1.3.1 Non-Truncated GLM & GAM

The non-truncated GLM and GAM modelling results are shown below on tables 3.1 and 3.2. The first table considers models fit to all 51 observations, whereas the second one only retains the 43 observations with the highest confidence level (confidence=1). Note that GAM is composed by *smooth* functions which have several coefficients. Therefore, there is no single parameter value associated with each smooth, and only the significance level of each *smooth* may be displayed.

Table 3.1: The best three candidate Poisson models (GLM and GAM) that explain the response variable "group size", along with the explanatory variables' coefficients (from GLM), and *smooth* significance level (from GAM), and the model's AIC value. The models were built considering all the 51 observations.

Model name	Model explanatory variables	GLM coeff.	GAM smooth p-value	GLM p-value	AIC value
A1	click duration	0.0125	0.1013	0.1013	142.2435
	number of hydrophones	−0.1122	0.0461*	0.0461*	
	click rate	0.0045	0.0049**	0.0049**	
A2	number of hydrophones	−0.0695	0.1499	0.1499	142.9390
	click rate	0.0039	0.0101*	0.0101*	
A3	click rate	0.0022	0.0160*	0.0160*	143.1803

Significance codes: 0.01 '**', 0.05, '*'

Table 3.2: The best three candidate Poisson models (GLM and GAM) that explain the response variable "group size", along with the explanatory variables' coefficients (from GLM), and *smooth* significance level (from GAM), and the model's AIC value. The models were built only considering the 43 groups with a confidence level of 1.

Model name	Model explanatory variables	GLM coeff.	GAM smooth p-value	GLM p-value	AIC value
B1	click duration	0.0045	0.5341	0.5341	122.0444
	whiskey hydrophones	−0.4718	0.0657	0.0657	
	click rate	0.0029	0.0055**	0.0055**	
B2	whiskey hydrophones	−0.4905	0.0546	0.0546	120.4265
	click rate	0.0032	0.0024**	0.0024**	
B3	click rate	0.0025	0.0095**	0.0095**	122.1139

Significance codes: 0.01 '**', 0.05, '*'

Both GLM and GAM methods hold the same variable choices, significance levels, and AIC results.

When merely considering the groups with a confidence level of 1, the importance of the variable "number of hydrophones" diminishes, being replaced by "whiskey hydrophones". The one model common in both tables is composed solely by the variable "click rate".

Furthermore, when attempting to fit a Negative Binomial GLM or GAM, a warning message would appear indicating a large θ parameter (the scale parameter of the Negative Binomial) that would not converge. For very large θ values the coefficient estimates are close to a Poisson distribution. (Klugman *et al.*, 2004), which is not surprising for our under dispersed data set.

3.1.3.2 Zero-truncated GLM & GAM

For the zero-truncated approach the analysis encountered an issue: when modelling with GAM, *R* would evoke a likelihood convergence error. This often occurs due to the sample size, which may not be large enough. Family functions from the *VGAM* package use the type of algorithm described in McCullagh (1980), where it is demonstrated that for sufficient large samples an unique maximum of the likelihood is guaranteed, whereas while modelling with smaller samples one may be confronted with convergence obstacles. This led us to continue the zero-truncated analysis solely with GLM. The results are presented in tables 3.3 and 3.4.

Table 3.3: The best three candidate Poisson models (zero-truncated GLM) that explain the response variable "group size", along with the explanatory variables' coefficients and the model's AIC value. The models were built considering all the 51 observations.

Model name	**Model explanatory variables**	**GLM coeff.**	**GLM p-value**	**AIC value**
C1	click duration	0.0103	0.1374	150.4935
	number of hydrophones	−0.0926	0.0713	
	click rate	0.0037	0.0113*	
C2	number of hydrophones	−0.0573	0.1931	150.7108
	click rate	0.0033	0.0205*	
C3	click rate	0.0018	0.0298*	150.5499

Significance codes: 0.01 '**', 0.05, '*'

Table 3.4: The best three candidate Poisson models (zero-truncated GLM) that explain the response variable "group size", along with the explanatory variables' coefficients and the model's AIC value. The models were built only considering the 43 groups with a confidence level of 1.

Model name	**Model explanatory variables**	**GLM coeff.**	**GLM p-value**	**AIC value**
D1	click duration	0.0037	0.5740	129.1026
	whiskey hydrophones	−0.3873	0.0960	
	click rate	0.0024	0.0124*	
D2	whiskey hydrophones	−0.4027	0.0821	127.4147
	click rate	0.002	0.0062**	
D3	click rate	0.0021	0.0196*	128.4403

Significance codes: 0.01 '**', 0.05, '*'

The models built under untruncated and zero-truncated analysis hold the same variables.

Considering models with similar AIC and according to the rule of parsimony, the simpler model should be picked. Also, the simpler model is the only one featured when modelling both with 51 or 43 observations. Therefore, in order to use the maximum amount of information possible, all the 51 observations will be considered. Such will prompt the study to continue with the model C3 , as seen on table 3.3.

3.1.4 Analysing the model

As mentioned before, it is important to analyse the model chosen for further inference, specially in terms of residuals and influential values.

3.1.4.1 Residuals

A Shapiro-Wilk test was applied to test the normality of the residuals. The results are as followed:

$\boldsymbol{H_0}$: The model's residuals are normally distributed

vs.

$\boldsymbol{H_1}$: The model's residuals are not normally distributed

$\boldsymbol{W} = 0.97532$
p-value $= 0.3626$

The null hypothesis is not rejected for any reasonable α level considered.

Figure 3.10 illustrates the residuals' behaviour against the fitted values, with a 95% confidence interval.

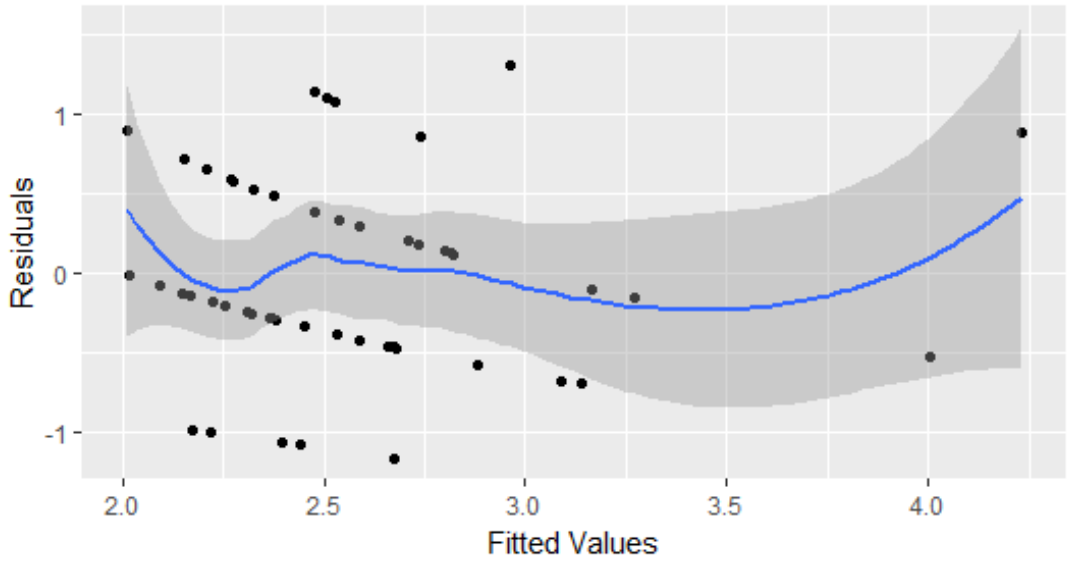


Figure 3.10: Fitted values and corresponding residuals, with a scatter plot smoother (grey area).

Additionally, a Pearson's product-moment correlation test between the fitted group size values and the residuals was employed:

$\boldsymbol{H_0}$: The correlation between the fitted values and their residuals is equal to zero.

vs.

$\boldsymbol{H_1}$: The correlation between the fitted values and their residuals is not equal to zero.

$R = -0.01525496$
p-value $= 0.9154$

The null hypothesis is not rejected, which indicates there is not a significant correlation between the model's fitted values and their residuals.

3.1.4.2 Hat values

The model may include some extreme hat values. Figure 3.11 provides a better visual insight.

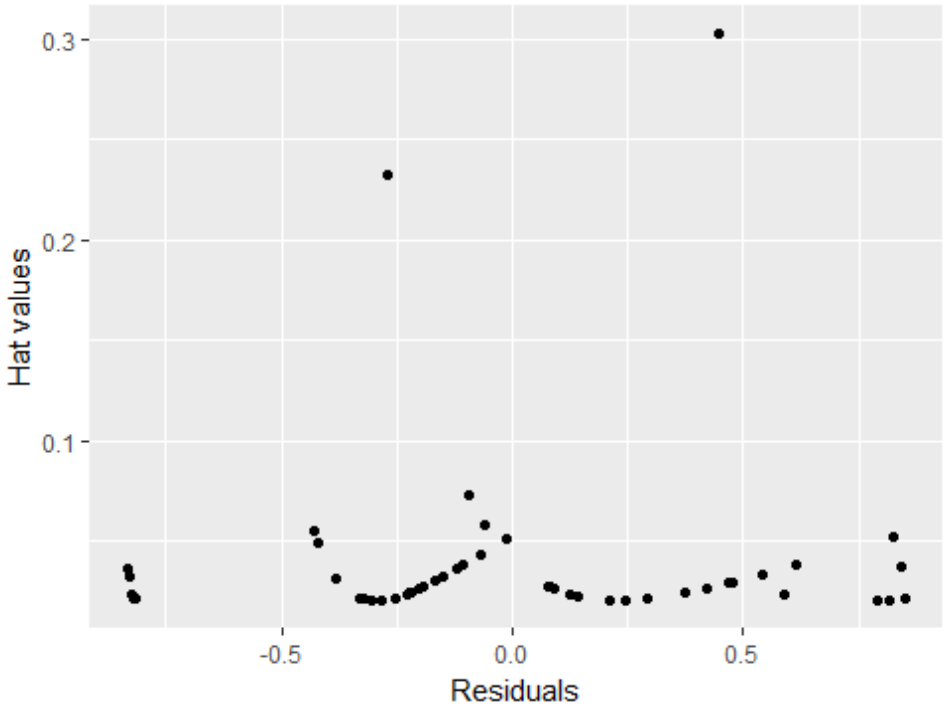


Figure 3.11: Model residuals and corresponding hat values

The extreme hat values correspond to observations 13 and 47. The later were removed in order to pinpoint their influence. The model was then rebuilt. Table 3.5 holds the results.

Table 3.5: The best three candidate Poisson models (zero-truncated GLM) that explain the response variable "group size", along with the explanatory variables' coefficients and the model's AIC value. The models were built considering all the 49 observations without extreme hat values.

Model name	**Model explanatory variables**	**GLM coeff.**	**GLM p-value**	**AIC value**
E1	click duration	0.0106	0.1831	142.7658
	number of hydrophones	−0.0974	0.0842	
	click rate	0.0038	0.0386*	
E2	number of hydrophones	−0.0594	0.1993	142.9766
	click rate	0.0031	0.0728	
E3	click rate	0.0014	0.2130	142.9038

Significance codes: 0.01 '**', 0.05, '*'

The same variables were selected as before, although it is noticeable the decrease in value regarding the "number of hydrophones" variable coefficient. Also, although the same variables were chosen, their significance level ($p-value$) decreased. This may be an indication that more data need to be collected to sustain a model.

Additionally, the observations without a confidence level of 1 were also removed and the model rebuilt. Table 3.6 holds the results.

Table 3.6: The best three candidate Poisson models (zero-truncated GLM) that explain the response variable "group size", along with the explanatory variables' coefficients and the model's AIC value. The models were built only considering the 41 observations without extreme hat values and with a confidence level of 1.

Model name	**Model explanatory variables**	**GLM coeff.**	**GLM p-value**	**AIC value**
F1	mean count	0.0003	0.4020	129.1026
	click rate	0.0018	0.2470	
	whiskey hydrophones	−0.2829	0.2990	
F2	click rate	0.0023	0.0841	127.4147
	whiskey hydrophones	−0.4038	0.0834	
F3	whiskey hydrophones	−0.3312	0.1480	128.4403

Significance codes: 0.01 '**', 0.05, '*'

It is noticeable that removing more observations in an already small dataset leads to a different selection of variables and a decrease on the variables' significance level.

Comparing all the modelling approaches, the model with the single explanatory variable "click rate" appears to be the best one. Once more, in furtherance of preserving the maximum amount of information available, the analysis proceeds with the model where all observations were considered (model C3).

3.2 Density Estimation Dataset

Similarly to the modelling dataset, the density estimation dataset was first subjected to a thorough exploratory data analysis.

3.2.1 Exploratory Analysis

To visualize possible click detection problems among the hydrophones, the variable "click count" was plotted against the 93 hydrophones. Figure 3.12 illustrates the click counts each hydrophone detected on the raw data.

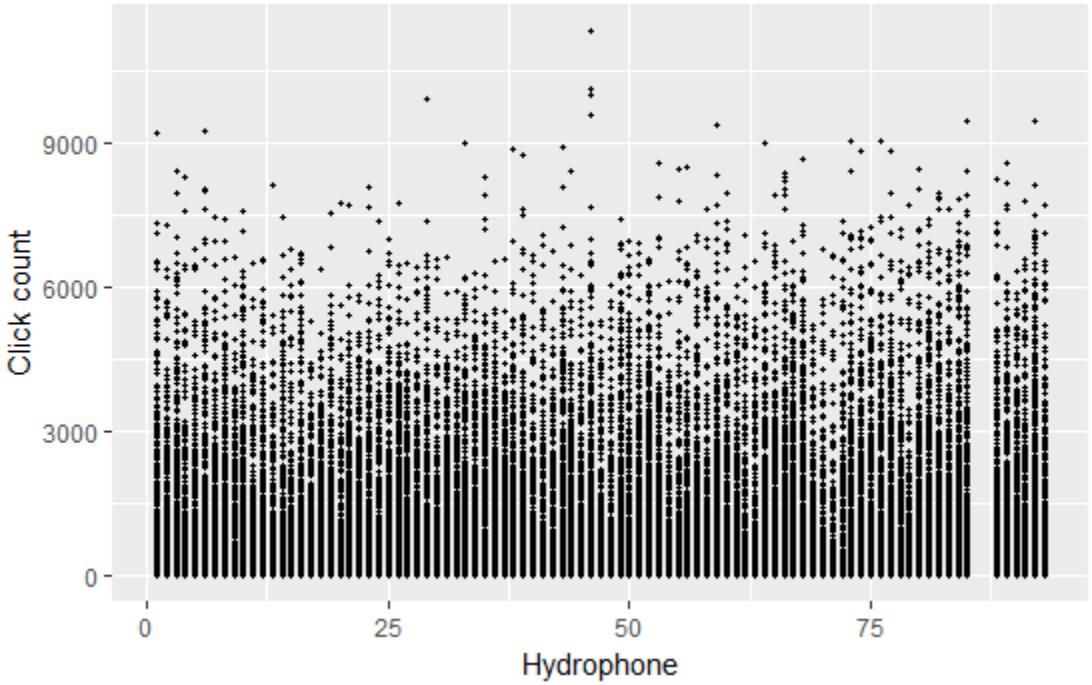


Figure 3.12: Click counts for each hydrophone (raw data).

One may notice that two of the hydrophones (86 and 87) have no detections at all, meaning there was probably an issue with them. The AUTEC was promptly notified of the situation.

3.2.2 Group Size Estimation

Group size was then estimated for every group. A mean group size value of 2.35 individuals per group was estimated for the three time periods. The figures bellow illustrate the mean group size per day, for the corresponding 3 periods: (1) figure 3.13, (2) figure 3.14, and (3) figure 3.15, with a group size mean value of approximately 2.36, 2.30, and 2.33, respectively.

Table 3.7 summarizes the group size values obtained for each time period.

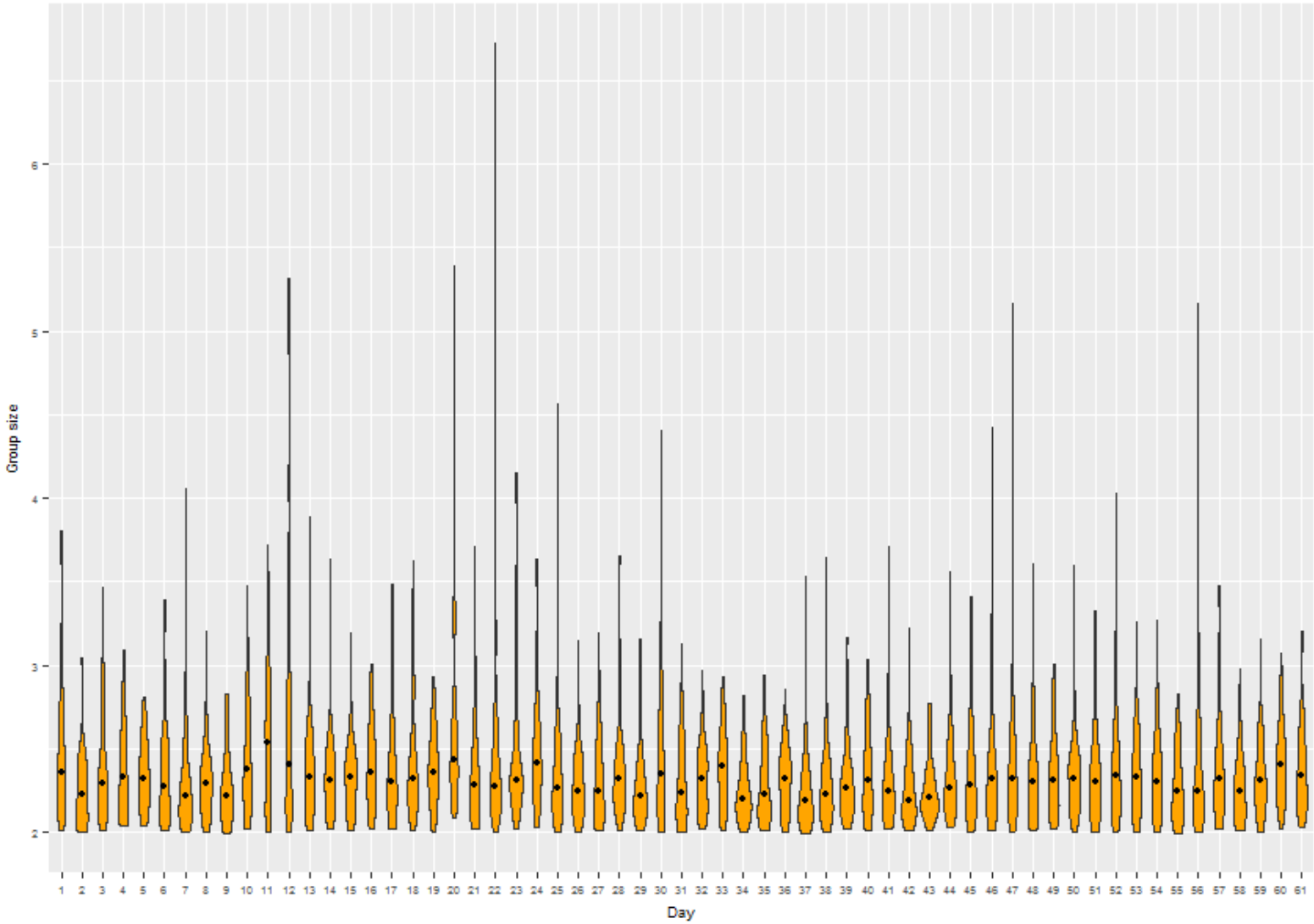


Figure 3.13: Group size estimations for each day (orange area), considering the first period (61 days). The orange area width is proportional to the number of estimated values for the group size. Each black dot inside every violin plot represents the median for the respective day.

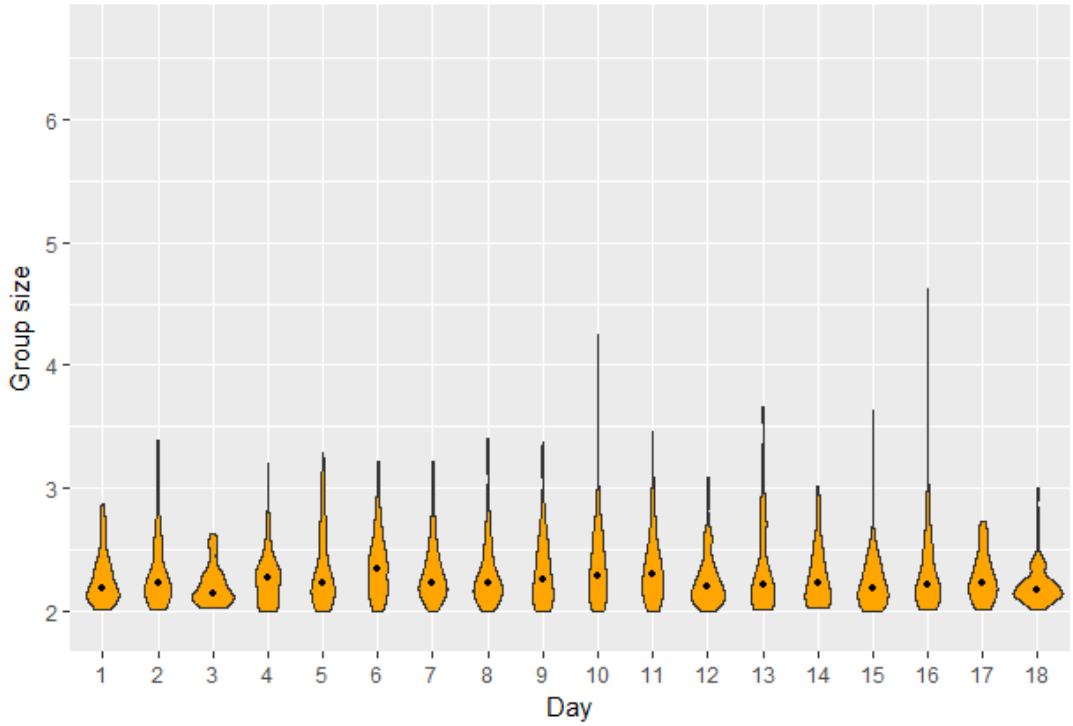


Figure 3.14: Group size estimations for each day (orange area), considering the second period (18 days). The orange area width is proportional to the number of estimated values for the group size. Each black dot inside every violin plot represents the median for the respective day.

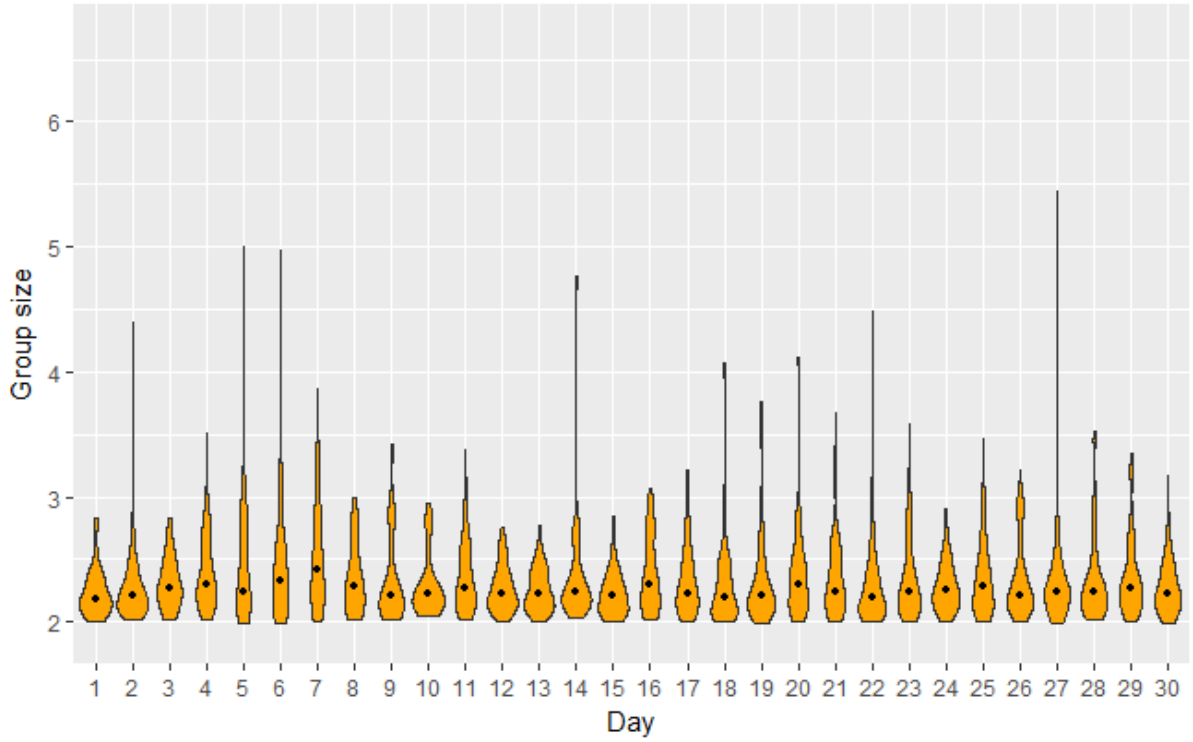


Figure 3.15: Group size estimations for each day (orange area), considering the third period (30 days). The orange area width is proportional to the number of estimated values for the group size. Each black dot inside every violin plot represents the median for the respective day.

Table 3.7: Group size estimation summary statistics for each of the three time periods considered.

	Period 1	**Period 2**	**Period 3**
Mean	2.3557	2.2976	2.3310
Standard Deviation	0.3144	0.2706	0.3117
Minimum	1.9898	1.9952	1.9914
Maximum	6.7230	4.6200	5.4425

3.2.3 Density Estimation

Finally, the density estimation (whales/1000 km^2) for each day was obtained. The overall mean density value estimation for all three periods is 15.91 whales/1000 km^2, with a mean value of 75.88 dives per day. The figures bellow illustrate the density estimation for the corresponding 3 time periods: (1) figure 3.16, (2) figure 3.17, and figure (3) 3.18, with respectively a density mean value of approximately 15.80, 16.45, and 15.81 whales/1000 km^2, and an average number of dives per day of 74.79, 79.94, and 75.67. The average number of dives per day for the three time periods is approximately 76.8.

Table 3.8 summarizes the density values obtained for each time period.

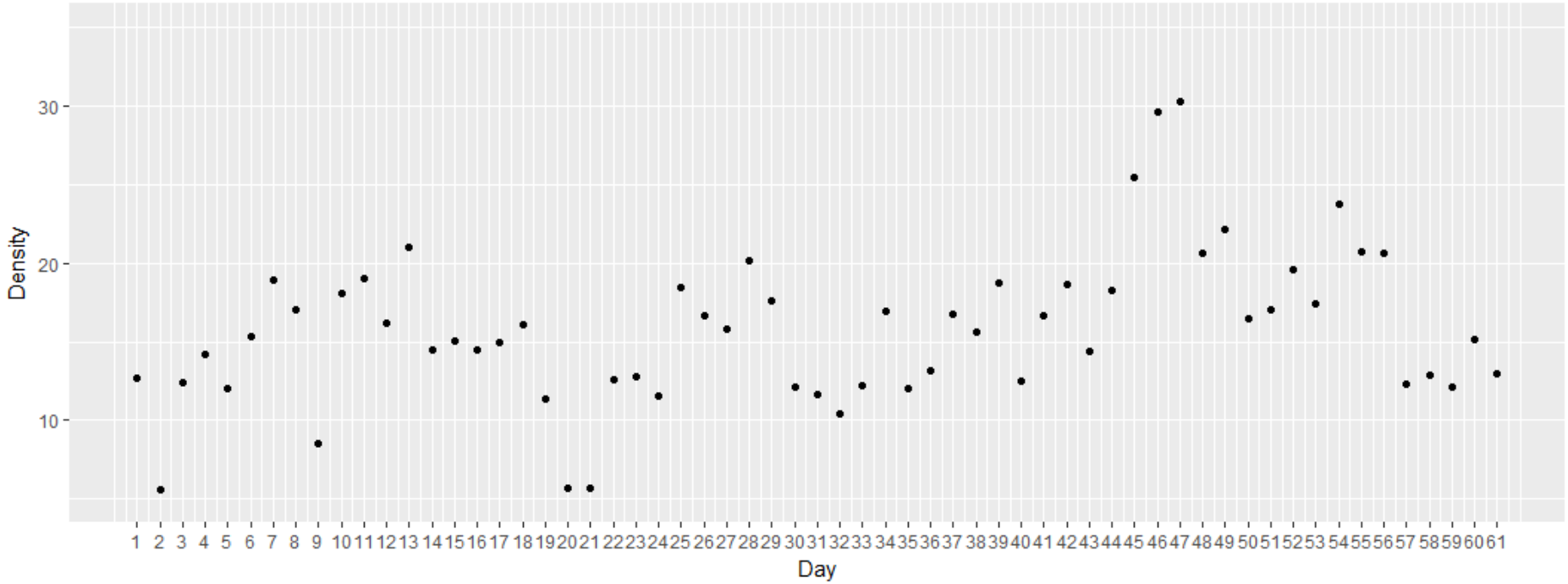


Figure 3.16: Density estimation for each day (whales/1000 km^2), considering the first period (61 days).

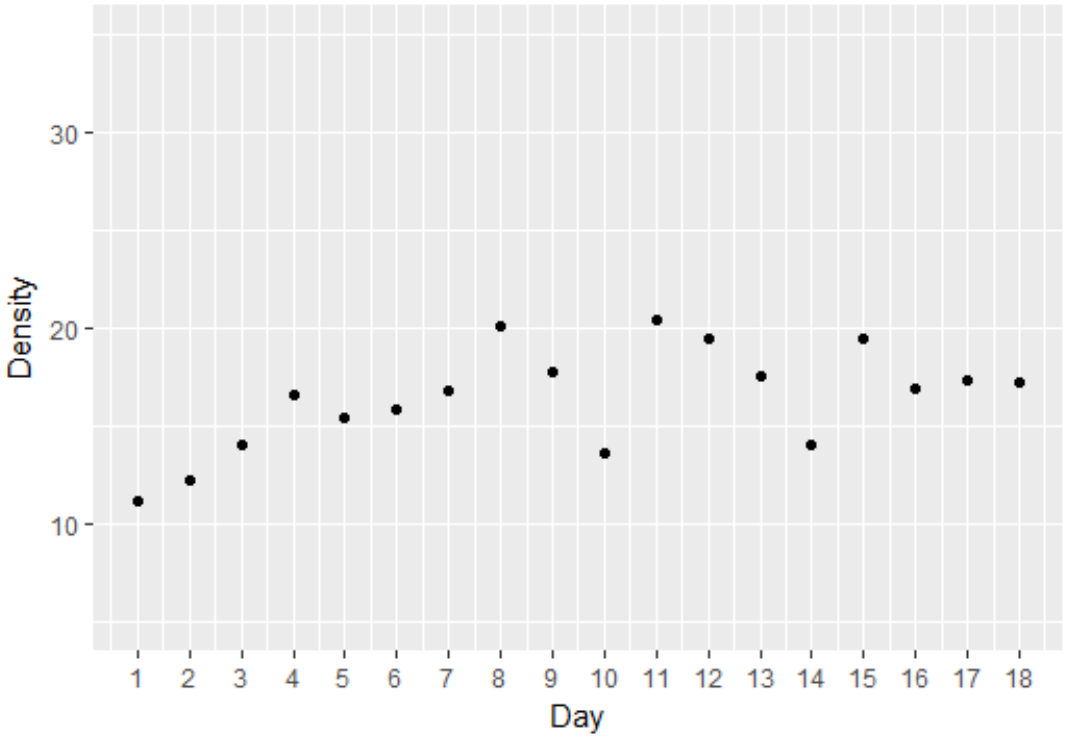


Figure 3.17: Density estimation for each day (whales/1000 km^2), considering the second period (18 days).

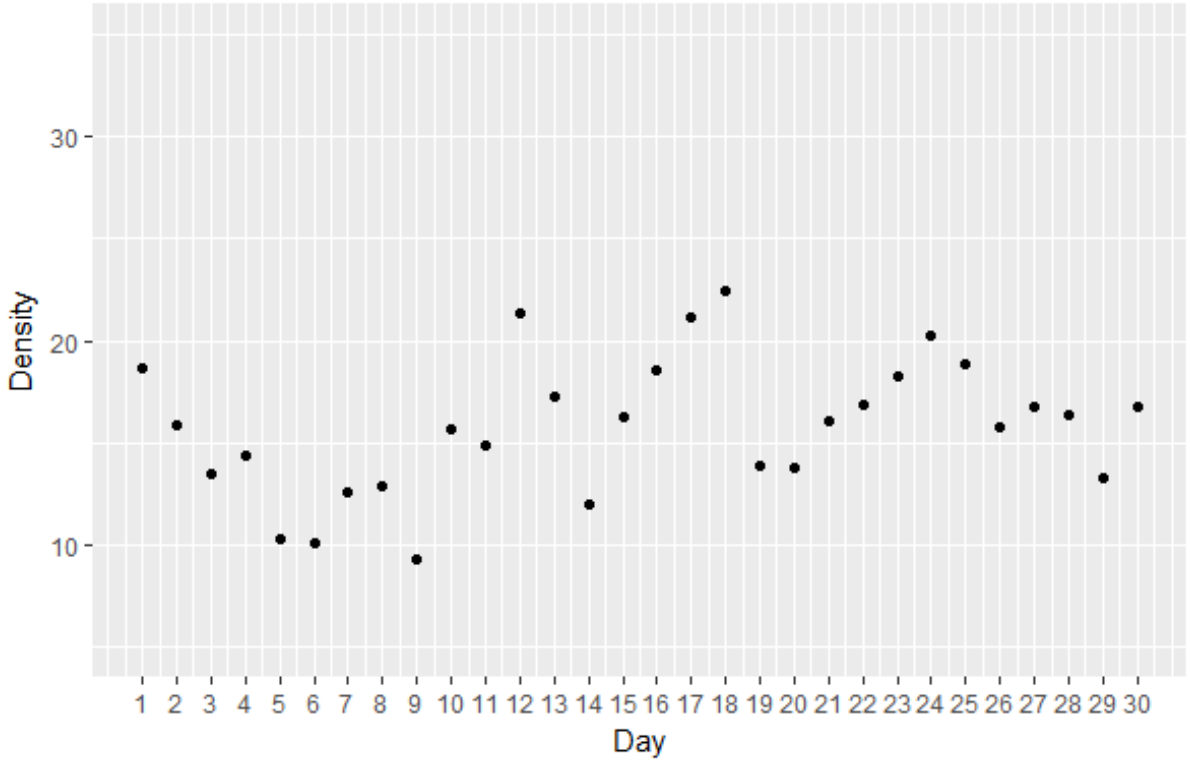


Figure 3.18: Density estimation for each day (whales/1000 km^2), considering the third period (30 days).

Table 3.8: Values regarding the density estimation for the three time periods.

	Period 1	**Period 2**	**Period 3**
Mean	15.7944	16.4674	15.8129
Standard Deviation	4.8373	2.6441	3.3323
Minimum	5.6368	11.1746	9.2678
Maximum	30.2686	20.4704	22.4915
Average # of groups/day	74.7869	79.9444	75.6667

3.2.4 Bootstrapping

To estimate the variance associated with the model for predicting group size, a bootstrap was implemented. The results are illustrated bellow.

Figure 3.19 illustrates the observed group size for each *Md* group, as well as the model's maximum likelihood fit line with a percentile interval of 95%.

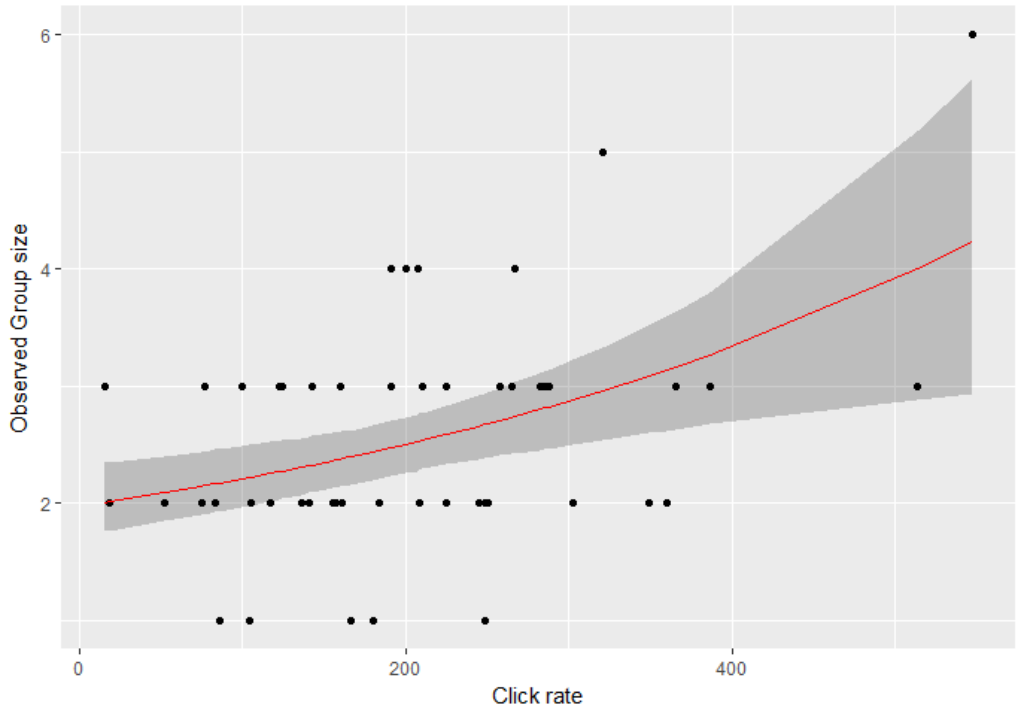


Figure 3.19: The observed group sizes and corresponding click rate (black dots), along with the model's maximum likelihood fit line (red line), and the model's bootstrap 95% percentile interval (grey area).

It was then necessary to propagate the variability of the model to the density estimation procedure, where a bootstrap exercise was also implemented. Figure 3.20 represents the 999 bootstraps (one for each colour), where the x-axis corresponds to the click rate variable and the y-axis to the group size.

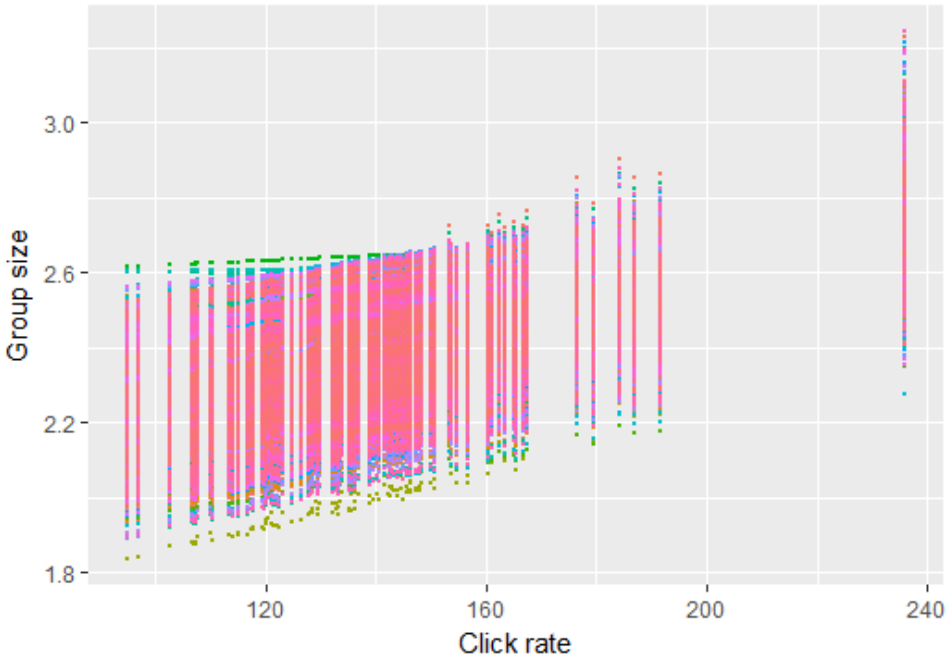


Figure 3.20: 999 group size bootstraps for the chosen model, for each click rate value. Although barely distinguishable, each colour represents a group size bootstrap for the corresponding click rate value.

Figure 3.21 is representing the estimated density per day for each bootstrap.

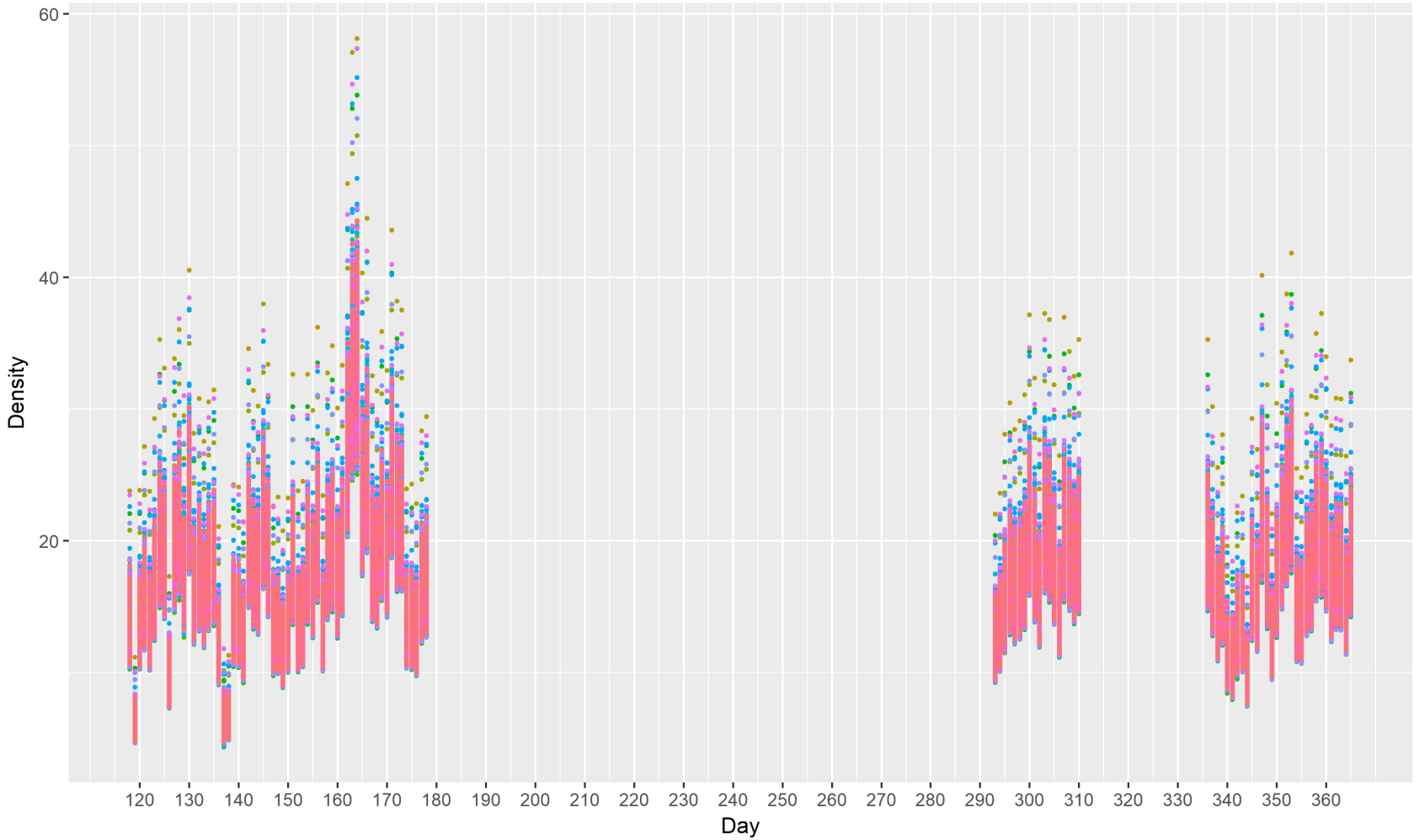


Figure 3.21: 999 model density bootstraps for each day, considering the three time periods. Although barely distinguishable, each colour represents a single bootstrap.

3.3 Comparison with Previous Results

One of the main goals of the present study is to compare the results with previous ones. Moretti *et al* (2010) compared estimated densities between time periods before, during and after sonar usage. Table 3.9 reproduces their results.

Table 3.9: Estimated abundance and density based on dive counting, with corresponding coefficient of variation (CV). Values in brackets after the estimates are 95% limits. Adapted from tables 1 and 3 in Moretti *et al.*, 2010.

Time period	**Abundance**	**Density (whales/1000 km^2)**	**Total # of groups**	**CV (%)**
Before sonar (65h prior to initial transmission)	22 (17-28)	16.99 (13.47 - 21.43)	194	11.89
During sonar (68.12h of transmission)	6 (4-8)	4.76 (3.78 - 6.01)	57	11.89
After sonar (65h after last transmission)	11 (8-14)	8.67 (6.87 - 10.94)	99	11.89
65h after sonar (43.23h)	32 (25-40)	24.76 (19.63 - 31.23)	188	11.89

Moretti et al (2010) used a single mean group size (s) of 2.62 animals/group, based on existing literature. Comparatively, Baird *et al* (2006) measured a Md group size value of 3.6 animals/group on the Big Island in Hawaii; while Claridge (2004) reported an average group of size of 4.1 animals/group on the Northern Bahamas.

According to table 3.9, the total number of groups detected per 24 hours would be approximately 71.63, 20.08, 36.55, and 104.37, respectively, for the four measurement periods.

Chapter 4

Discussion

News regarding catastrophic wildlife declines are common. In fact, these reports are becoming more prominent as we are dealing first-hand with real environmental consequences. When digesting that information, a question we often guiltily ask ourselves is "what can I do to help?". That feeling of culpability often vanishes after a while. Fortunately, in other occasions, such spark will not vanish, ending up being the foundation of something meaningful. Scientists are often the ones keeping their spark alive, presenting and innovating methods aiming better conservancy policies. An example lies with Passive Acoustic Monitoring, a field with a great potential when it comes to study wild marine populations. PAM presents us the chance of accurately estimating wild animal population size and density. Since it relies on acoustic footprints, its performance is not compromised by dark environments. Also, each species produce distinctive sounds, and because acoustic signs are often detectable at greater distances, it often makes them a more reliable source of information than visual cues. Due to its distinctive and innovative characteristics, PAM may become an important tool for the future of Ecology and Conservation.

Since *Md* is a species known to produce echolocation sounds, but hard to detect visually, PAM may be an efficient solution to collect data about it. Presently, *Md* does not have an established conservation status due to lack of information (iucnredlist.org), which may be adjusted in a near future as we obtain additional information on the species. The present work aims to pave that change by contributing with this analysis of *Md* data.

The density estimation formula (2.53) has two random components associated with it, the group size and the dive rate. The proposed method improves on the previous approach of Moretti *et al* (2010) by (1) allowing the estimation of a group size for each group, and hence (2) allowing the estimation of a mean group size for each period of interest, and therefore (3) allowing to relax the implicit assumption that group size is constant over time and space. However, the same problem still applies to the dive rate, which is taken from the literature, based on a small sample of tagged animals (Moretti *et al*, 2010), and assumed constant over time. It is possible that differences of dive rates are larger over time and space than differences in group sizes, meaning that while a useful step in obtaining more reliable estimates, dealing with variation in group size might fall short from being enough to get reliable density estimates from dive counting methods. This means that additional studies looking at dive rates from beaked whales, investigating how these might change in time (e.g., seasonally) and in space (e.g., being or not depth dependent), are fundamental to understand the reliability of dive counting methods.

4.1 Underlying Assumptions

In this work we assumed no false positives and no missed detections, nonetheless there is some evidence that a small number of these might occur (D. Moretti 2016, personal communication).

It seems important to refer that a study looking at these two assumptions would be much welcome. Depending on the magnitude of each of these phenomena the density estimates under these assumptions might be underestimating or over estimating density. Given that these phenomena occur at a small scale and the induced biases will have different signs, with false positives overestimating abundance, but missed detections underestimating abundance. Therefore, despite perhaps more elegant and thorough analysis might be conducted while accounting for them, we do not anticipate major changes once these factors are actually incorporated.

4.2 Conclusions

We conclude that, based on the acoustic footprint of groups detected on AUTEC hydrophones, the variable "click rate" appears to be the best descriptor of group size. However, when it comes to modelling, it is noticeable that more observations may be needed, as a small data set will never allow a complex model to be a parsimonious choice. Therefore, it is possible that with additional data more complex models might prove useful to describe group size from the group's acoustical footprint. Although the model composed solely by the "click rate" variable was always among the models' top 3, the variable "number of hydrophones" was replaced by "whiskey hydrophones" on the remaining two models when only considering the groups with a confidence level of 1, which may indicate difficulties when choosing between variables.

Moreover, the variable "number of hydrophones" held a negative coefficient, which is not logically or physically plausible since a bigger group would trigger more hydrophones, and not the other way around. This could be due to several groups being detected over Whiskey hydrophones: it would be possible for a smaller group to trigger a higher number of hydrophones if it happened to dive near Whiskey hydrophones. That is enough to introduce confounding and the consequent difficulty in getting a reliable model. Therefore, a different way of incorporating the information from these hydrophones in the analysis might be preferable, which should be investigated in the continuation of this study.

In terms of the estimated group sizes and densities, the results presented here are well in line with those from previous studies (e.g., Moretti *et al.*, 2010). Nonetheless, the present work obtained a smaller mean group size estimate than the one from literature (2.35 *vs* 2.62). The difference is higher when considering the values reported in Claridge (2004) and Baird *et al* (2006), (4.1 and 3.6 respectively). The present work estimates a mean density of 15.91 across all time periods. It is presumed that no major sonar activity took place during the echolocation clicks detection, although small activities can not be ruled out, both during this study, before or after. Although interpreting the results might be easier knowing times of sonar emission, information about sonar activity at the AUTEC range is extremely sensitive and therefore unlikely to be available to us.

It is important to acknowledge that the density values reported by Moretti *et al.* (2010) are 16.99, 4.76, 8.67 and 24.76 were calculated considering much smaller time periods (65h, 68.13h, 65h and 43.23h respectively), for which the impact from the sonar activity may have become more perceptible. Therefore, localized sonar activity occurring for the present three time periods (1464h, 432h and 720h, respectively) would probably not gravely reflect on the total estimated densities, although it would on the daily estimates. This may indicate that reactions are fast,

and considering a larger scale the effect of the sonar activity on *Md* density is barely perceptible. It is clear that some density fluctuation occurs over time. This fluctuation is also apparent when visualizing the density bootstrap figure (figure 3.21). This indicates that the estimation of the group size is an important factor when estimating density.

Additionally, the number of groups detected per day appears to be consistent among the three time periods (74.8, 79.9 and 75.7, respectively), with a global mean of 75.9 groups/24h. Moretti *et al.* (2010) results for before, during and after sonar activities seem to exhibit larger differences (97.1, 32.8 and 24.0 groups/24h, respectively). That is not surprising due to shorter time (65h, 68h and 365h, respectively), and the fact that sonar activities were occurring, presumably driving off *Md* individuals from the AUTEC range.

One advantage of the proposed method is to be able to provide a mean group size estimate for any time period that one might consider. These differences would be averaged out when making comparisons across time points having to share the same mean estimate obtained from the literature. This kind of data could be used in itself to derive spatio-temporal models of group size at AUTEC.

These type of studies aim conservation purposes. It is important to acknowledge their importance and their necessity to be constantly updated. Describing the group size over time may contribute to better understand *Md* species habits, which leads to better conservation methods, since a species' density fluctuation over time may be due to several external factors, which may also include human disturbance. Those are important to target, in order to minimize human-made impact. At the current pace that species are being affected by habitat deterioration, model improvements are vital when it comes to study natural populations, as they provide more accurate information which will help to provide decisions based on evidence leading to an effective management of ecosystems.

Future work on this area are will include gathering more group size data for visually confirmed groups; a different approach for the inclusion of the differential detection capabilities of Whiskey hydrophones, and to extend the density estimation by incorporating false positives and non-detected groups.

4.3 Acquired Competencies

The work presented here was a partial requirement for obtaining an MSc in Biostatitsics. It is worth to list explicitly the set of tools and statistical competences that it allowed me to develop and become familiar with:

1. Manipulate large datasets;
2. Perform a thorough exploratory data analysis;
3. Resort to different correlation measures between different types of variables;
4. Recognize when variables interact or correlate, and be able to deal with that issue;
5. Properly apply and distinguish different statistical tests;
6. Being able to implement LM, GLM, GAM, and zero-truncated models;
7. Deal with data from different distributions, especially Poisson and Negative Binomial;

8. Build and chose between models resorting to different approaches;
9. Employ different goodness-of-fit tools;
10. Implement both non-parametric and parametric bootstraps to estimate variances when analytical expressions are not available;
11. Able to work and compile large documents in LaTeX;
12. Able to work with several R packages (to name a few: *ggplot2, cowplot, vioplot, VGAM, mcgv, leaps, boot*);
13. Develop coding and programming capabilities;
14. Adhere to the ideas of reproducible research using dynamic reports (Appendix A).

4.4 Final Remarks

Statistical procedures can be used to obtain practical answers to biological questions. In this work we have estimated beaked whale density at AUTEC during a 4 month period. Additionally, we have obtained the associated precision measures without which density estimates would be meaningless. While doing so we have identified a number of follow up research questions and gained a number of statistical competencies. By providing the data and the code used via a dynamic report, we make the research process completely transparent, allowing readers to implement themselves the analysis presented, adhering to the uprising paradigm of reproducible research (Peng, 2011).

References

♢ **Altman, D. G.** (1991). *Practical Statistics for Medical Research.* London: Chapman and Hall.

♢ **Baird, R.W., Webster D.L., McSweeney D.J., Ligon A.D., Schorr G.S.,** & **J. Barlow** (2006). Diving behaviour of Cuvier's (*Ziphius cavirostris*) and Blainville's (*Mesoplodon densirostris*) beaked whales in Hawai'i. *Canadian Journal of Zoology*, 84, 1120-1128.

♢ **Baird, R.W., Webster D.L.,Schorr G.S., McSweeney D.J.,** & **Barlow J.** 2008. Diel variation in beaked whale diving behavior. *Marine Mammal Science* 24, 630-642.

♢ **Barlow, J.** (1999). Trackline detection probability for long-diving whales. In G. W. Garner, S. C. Amstrup, J. L. Laake, B. J. F. Manley, L. L. McDonald & D. G. Robertson (Eds.) *Marine Mammal Survey and Assessment Methods* (pp. 209-221). Netherlands: A.A. Balkema Publishers.

♢ **Burnham, K. P., Anderson, D. R.** (2002). *Model Selection and Multimodel Inference A Practical Information-Theoretic Approach.* New York: Springer-Verlag.

♢ **Cameron, A. C., Trivedi, P. K.** (1998). *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

♢ **Chatterjee, S.** & **Hadi, A.** (1986). Influential observations. *Statistical Science*, 1, 379-392.

♢ **Chernick, M. R.,** & **LaBudde, R. A.** (2011). *An Introduction to Bootstrap Methods with Applications to R.* New Jersey: Wiley.

♢ **Claridge, D. E.** (2004). *Fine-scale distribution and habitat selection of beaked whales* (MSc thesis). University of Aberdeen, Scotland.

♢ **Conover, W. J.** (1999). *Practical Nonparametric Statistics.* New York: Wiley.

♢ **Cramer, H.** (1946). *Mathematical Methods of Statistics.* New Jersey: Princeton University Press.

♢ **Cramer, D.** (1998). *Fundamental Statistics for Social Research.* London: Routledge.

♢ **DiMarzio, N., Moretti, D., Ward ,D., Morrissey, R., Jarvis, S., Izzi, A., Johnson, M., Tyack,P.,** & **Hansen, A.** (2008). Passive acoustic measurement of dive vocal behavior and group size of Blainville's beaked whale (*Mesoplodon densirostris*) in the tongue of the ocean (TOTO). *Canadian Acoustics*, 36, 166-173.

♢ **Draper, N.,** & **Smith, H.** (1981). *Applied Regression Analysis.* New York: Wiley.

♢ **Efron, B.** & **Tibshirani, R.** (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

♢ **Fabozzi, F. J., Focardi, S. M., Rachevand, S. T.,** & **Arshanapalli, B. G.** (2014). *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications.* New Jersey: Wiley.

♢ **Frome, E. L.** (1982). Algorithm AS 171: Fisher's Exact Variance Test for the Poisson Distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31, 67-71.

♢ **Greenwood, P. E.,** & **Nikulin, M. S.** (1996). *A Guide to Chi-Squared Testing.* New Jersey: Wiley.

♢ **Hair, J.F. Jr., Anderson, R.E., Tatham, R.L.,** & **Black, W.C.** (1998). *Multivariate Data Analysis.* New Jersey: Prentice Hall.

♢ **Hardin, J. W., Hilbe, J. M.** (2007). *Generalized Linear Models and Extensions.* Texas: Stata Corp.

♢ **Hastie, T. J.,** & **Tibshirani R. J.** (1986). Generalized additive models. *Statistical Science*, 1, 295–318.

♢ **Hastie, T. J.,** & **Tibshirani, R. J.** (1990). *Generalized Additive Models.* New York: Chapman and Hall.

♢ **Jefferson, T. A., Webber, M. A.,** & **Pitman, R. L.** (2008). *Marine Mammals of the World, A Comprehensive Guide to their Identification.* Amsterdam: Elsevier.

♢ **Johnson, M., Madsen, P. T., Zimmer, W. M. X., Aguilar de Soto, N.,** & **Tyack, P. L.** (2006). Foraging Blainville's beaked whales (*Mesoplodon densirostris*) produce distinct click types matched to different phases of echolocation. *The Journal of Experimental Biology*, 209, 5038-5050.

♢ **Klugman, S.A., Panjer, H. H.,** & **Wilmot, G. E.** (2004). *Loss Models, From Data to Decisions.* New York: Wiley

♢ **Kutner, M. H., Nachtsheim, C. J.,** & **Neter, J.** (2004). *Applied Linear Regression Models.* Illinois: McGraw-Hill Irwin.

♢ **Leatherwood, S.,** & **Reeves, R.R.** (1983). *The Sierra Club Handbook of Whales and Dolphins.* San Francisco: Sierra Club Books.

♢ **MacLeod, C. D.** (1998). Intraspecific scarring in odontocete cetaceans: an indicator of male 'quality' in aggressive social interactions? *Journal of Zoology*, 244, 71 – 77.

♢ **Manly B. F. J.** (2006) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Boca Raton, FL: Chapman and Hall/CRC.

♢ **Marques, T. A., Thomas, L., Ward, J., DiMarzio, N.,** &**Tyack, P. L.** (2009). Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainville's beaked whales. *The Journal of the Acoustical Society of America*, 125, 1982-1994.

♢ **Marques T., Shaffer J.,** & **Thomas L.** (2013). Modelling group size as a function of autogrouper outputs. University of St Andrews, Scotland.

♢ **Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D.,** & **Tyack, P. L.** (2013). Estimating animal population density using passive acoustics. *Biological Reviews*, 88, 287-309.

♢ **McCann, C.** (1963). Occurrence of Blainville's beaked-whale *Mesoplodon densirostris* (Blainville) in the Indian ocean. *Journal of the Bombay Natural History Society*, 60, 727-731.

♢ **McCullagh, P.** (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, 109–142.

♢ **McCullagh, P.** & **Nelder, J.A.** (1989) *Generalized Linear Models*. Washington, DC: Chapman and Hall/CRC.

♢ **Mead, J.G.** (1989). *Handbook of Marine Mammals*. London: Academic Press.

♢ **Montgomery, D. C.,** & **Peck, E. A.** (1992). *Introduction to Linear Regression*. New York: John Wiley.

♢ **Moretti, D., Marques, T., Thomas, L., DiMarzio, N., Dilley, A., Morrissey, R., McCarthy, E., Ward, J.,** & **Jarvis, S.** (2010). A dive counting density estimation method for Blainville's beaked whale (*Mesoplodon densirostris*) using a bottom-mounted hydrophone field as applied to a Mid-Frequency Active (MFA) sonar operation. *Applied Acoustics*, 71, 1036-1042.

♢ **Pastene, L. A., K. Numachi, M. Jofre, M. Acevedo,** & **G. Joyce.** (1990). First record of the Blainville's beaked whale, *Mesoplodon densirostris* Blainville, 1817 (Cetacea, Ziphiidae) in the eastern South Pacific. *Marine Mammal Science*, 6, 82-84.

♢ **Peng, R. D.**. (2011). Reproducible Research in Computational Science, *Science* 334, 1226-1227.

♢ **R Core Team**. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.

♢ **Rencher, A. C.,** & **Schaalje, G.B.** (2008). *Linear Models in Statistics*. New Jersey: Wiley.

♢ **Sarhan, A. E.** & **Greenberg,B. G.** (1956). Estimation of location and scale parameters by order statistics from singly and double censored samples. Part I. *The Annals of Mathematical Statistics*, 27, 427-51.

♢ **Shaffer, J., Moretti, D., Jarvis, S., Tyack,** & **P., Johnson, M.** (2013). Effective beam pattern of the Blainville's beaked whale (*Mesoplodon densirostris*) and implications for passive acoustic monitoring. *The Journal of the Acoustical Society of America*, 133, 1770–1784.

♢ **Shaffer, J.,** & **Baggenstoss, P.** *Beaked Whale Group Deep Dive Behavior from Passive Acoustic Monitoring*. Retrieved from: https://www.onr.navy.mil/reports/FY15/mbshaffe.pdf

♢ **Shapiro, S. S.,** & **Wilk, M. B.** (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

♢ **Sheskin, D.** (2011). *Handbook of Parametric and Non Parametric Statistical Procedure*. Boca Raton, FL: CRC Press.

♢ **Yee, T. W., Stoklosa, J.,** & **Huggins, R. M.** (2015). The VGAM Package for Capture-Recapture Data Using the Conditional Likelihood. Journal of Statistical Software, 65, 1-33. URL http://www.jstatsoft.org/v65/i05/.

♢ **Tyack, P. L., Johnson M., Soto, N. A., Sturlese, A., Madsen P. T.** (2006). Extreme diving of beaked whales. *The Journal of Experimental Biology*, 209, 4238-4253.

♢ **Wayne, D. W.,** & **Cross, C. L.** (2013). *Biostatistics: A Foundation for Analysis in the Health Sciences*. Toronto, ON: Wiley.

♢ **Wickham, H.** (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

♢ **Wood, S.N.** (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

♢ **Yates, D., Moore, Moore, D.,** & **McCabe, G.** (1999). *The Practice of Statistics*. New York: W.H. Freeman.

♢ **Yee, T.** (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York: Springer.

♢ **Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A.,** & **Smith, G. M.** (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

♢ **Zuur, A. F.,** & **Ieno, E. N.** (2016). *Beginner's Guide to Zero-Inflated Models with R*. Newburgh, NY: Highland Statistics Ltd.

Appendix A

1 Introduction

This document serves the purpose of aiding the visualization and interpretation of the dissertation's content by including the code itself.

The main goal is to estimate the density of *Mesoplodon densirostris* as described in Moretti *et al* (2010). In order to do so, the group size needs to be estimated via acoustic footprint variables, such as the number of clicks, the click duration, the click rate, among others. These clicks are echolocations the species produces to communicate and capture prey, and are collected by a field of hydrophones located at the AUTEC.

The analysis starts with a modelling where the group size was confirmed for 51 groups, with which the modelling process takes place; followed by the density estimation dataset where the chosen model is employed. The density estimation datset was collected by the AUTEC hydrophones over a course of 4 months, where the group size was not confirmed for any group. With this work we hope to contribute to the creation of an automatized method to estimate *Md* density over time.

The analysis will resort to a zero-truncated GLM approach, since a group with zero elements is not a group at all. Also, in order to compare results, non-truncated GLM will also be employed. But in this later case the response variable needs to be transformed so no zero values are predicted. GAM were also applied, but won't be considered in this report, since zero-truncated GAM did not converge for this particular dataset.

Afterwards, in order to incorporate the variability from the first dataset and from the dive rate value (Moretti et al, 2010) into the model, a bootstrap was applied.

Note that some of the figures previously included in this work won't show up in this document as they occupy a large amount space.

2 The Modelling Dataset

The modelling dataset is composed by two separate tables which are then merged. The first table acknowledges the detailed information for each group. It considers the number of times the groups were detected, how many hydrophones were involved, the number of echolocation clicks, and the time period. The second table holds the number of whales confirmed in each group, as well as its confidence level (it goes from 1 to 3, where 1 corresponds to "more certain", and 3 to "less certain").

2.1 Reading and preparing the data

Both tables refer to 51 whale groups with a confirmed group size, whose echolocations were detected by the hydrophones.

```
#Reading the first table
dados.mod <- read.table("dadosi2.txt", header=T)
nomes <- names(dados.mod) <- c("gID","month","day","year","hyd",
                  "count","shou","smin","ssec","ehou","emin","esec")

#Reading the second table
dados.mod2 <- read.table("dadosi.txt",header=F)
names(dados.mod2) <- c("gID","date","cs","conf")

#Merging the tables' information
```

```
gdatac <- dados.mod[1,]
n <- nrow(dados.mod)
cr = 2
for (i in 2:n) {
    #if it is the same hydrophone, same group
  if (dados.mod$gID[i] == dados.mod$gID[i-1] & dados.mod$hyd[i] == dados.mod$hyd[i-1]) {
    #adding the click count to the previous row
    gdatac[cr-1,6] = gdatac[cr-1,6]+dados.mod[i,6]
    #get the end of the recorded vocal group in that phone
    gdatac[cr-1,10:12] = dados.mod[i,10:12]
  } else {
    gdatac[cr,] = dados.mod[i,]
    cr = cr+1
  }
}
```

Additional information was then included, such as the time duration at each hydrophone, and whether or not the hydrophone was Whiskey or Bi/Uni-directional.

```
#Getting the duration at each hyd
gdatac$date <- with(gdatac,paste(month,day,year,sep="-"))
gdatac$stime <- with(gdatac,paste(shou,smin,ssec,sep=":"))
gdatac$etime <- with(gdatac,paste(ehou,emin,esec,sep=":"))
gdatac$stime <- with(gdatac,strptime(paste(date,stime),format="%m-%d-%Y %H:%M:%S"))
gdatac$etime <- with(gdatac,strptime(paste(date,etime),format="%m-%d-%Y %H:%M:%S"))
gdatac$hyddur <- with(gdatac,difftime(time1=etime,time2=stime,units=c("mins")))

#Recoding Whiskey and Bi/Uni hydrophones
gdatac$whiskey <- gdatac$hyd
gdatac$direction <- gdatac$hyd
gdatac$whiskey <- car::recode(gdatac$whiskey, "1:14=TRUE")
gdatac$whiskey <- car::recode(gdatac$whiskey, "15:93=FALSE")
gdatac$direction <- car::recode(gdatac$direction,
"1=0;15=1;20=1;30=1;41=1;42=1;45=1;56=1;58=1;61=1;69=1;72=1;75=1;78=1;88=1;91=1;93=1")
gdatac$direction <- car::recode(gdatac$direction, "2:93=0")
```

Some other information was added: the maximum count and the number of clicks detected by each hydrophone, the mean count of clicks for each hydrophone, and the number of hydrophones involved with each group.

```
#Maximum count at a hydrophone
maxcount <- with(gdatac,tapply(X = count, INDEX = gID, FUN = max))
#Total count at all hydrophones
nclicks <- with(gdatac,tapply(X = count, INDEX = gID, FUN = sum))
#Mean count at all hydrophones
meancount <- with(gdatac,tapply(X = count, INDEX = gID, FUN = mean))
#Number of hydrophones involved
nhyd <- with(gdatac,tapply(X = hyd, INDEX = gID, FUN = length))
```

Furthermore, two different functions were built in order to differentiate the groups which were detected by at least a Whiskey or a by Bi-directional hydrophone.

```
#Whiskey hydrophone
has.wisk <- function(dados){
  x = 0
  if (sum(dados%in%1:14)>0) x = 1
  return(x)
```

```
}
wisk <- with(gdatac,tapply(X = hyd, INDEX = gID, FUN = has.wisk))

#Bi-directional hydrophone
direc <- unique(gdatac$gID)
i = 1
for (i in direc){
  x = 0
  g0 <- gdatac[gdatac$gID==i,]
  if (sum((g0$hyd%in%15|g0$hyd%in%20|g0$hyd%in%30|g0$hyd%in%41|
           g0$hyd%in%42|g0$hyd%in%45|g0$hyd%in%56|g0$hyd%in%58|
           g0$hyd%in%61|g0$hyd%in%69|g0$hyd%in%72|g0$hyd%in%75|
           g0$hyd%in%78|g0$hyd%in%88|g0$hyd%in%91|g0$hyd%in%93))>=1)
  x = 1
  direc[i] = x
}
```

A further variable was added: the vocal click duration for each group.

```
#Getting each group vocal period duration
cdur <- numeric(max(gdatac$gID))
for (i in 1:max(gdatac$gID)) {
  temp1 <- gdatac$etime[gdatac$gID==i]
  ge <- max(temp1)
  temp2 <- gdatac$stime[gdatac$gID==i]
  gs <- min(temp2)
  cdur[i] <- difftime(time1 = ge,time2 = gs,units = c("mins"))
}
```

Although some of the previous information is not accounted as acoustic footprint, and will therefore not be included in the modelling process, it is always relevant to visualize and compare differences among the groups.

Finally, the data is bundled up in a single data frame.

```
#Modelling data frame
d4reg <- data.frame(gID = dados.mod2$gID, cs = dados.mod2$cs, conf = dados.mod2$conf,
                    maxcount = as.numeric(maxcount), meancount = as.numeric(meancount),
                    nhyd = as.numeric(nhyd), cdur = as.numeric(cdur),
                    nclicks = as.numeric(nclicks), crate = as.numeric(nclicks/cdur),
                    wisk = as.numeric(wisk), direction = as.numeric(direc))
```

A new variable, click rate (**crate**), was included in the data frame. Click rate is obtained simply by dividing the corresponding group number of clicks by the click duration.

Since the modelling process will resort not only to a zero-truncated approach GLM, a transformation of the response variable is needed so no zero values are predicted.

```
d4reg$cs0 <- d4reg$cs-1
```

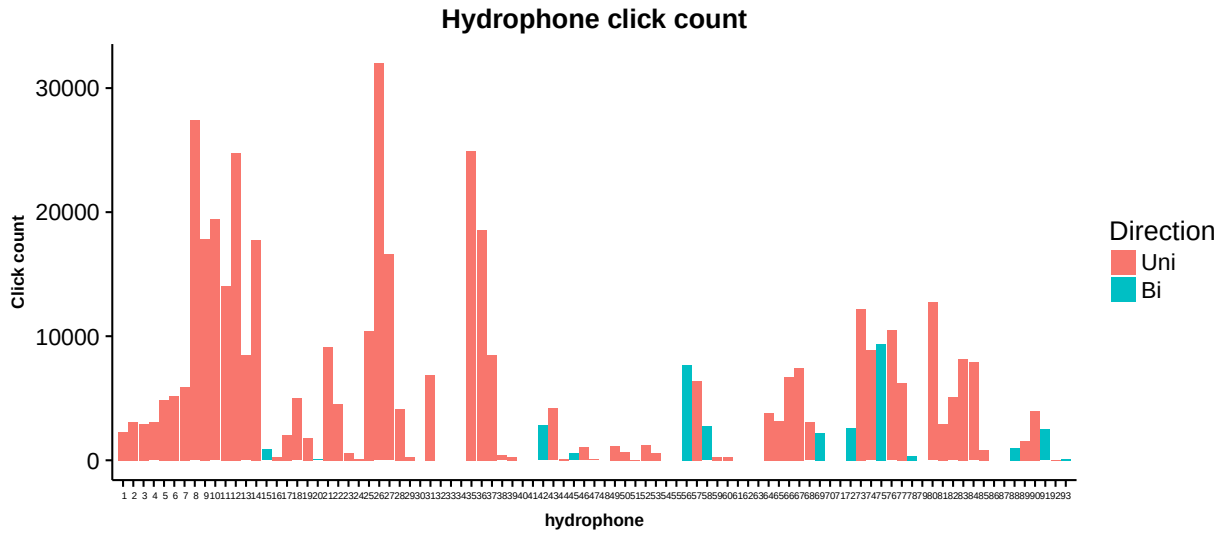
Also, it is important to factorize the Whiskey and Direction variables.

```
d4reg$wisk <- factor(d4reg$wisk)
d4reg$direction <- factor(d4reg$direction)
```

2.2 Exploratory analysis

The figure bellow reveals the number of clicks each hydrophone counted. It's important to have in account the hydrophones numbered 1-14 are Whiskey, and the blue ones are Bi-directional.

```
#Histogram, counts per hyd
ggplot(gdatac, aes(x = hyd, y = count,colour,fill = factor(direction,
             labels = c("Uni","Bi")))) + geom_bar(stat="identity") +
  scale_x_discrete(limits=c(1:93)) + ggtitle ("Hydrophone click count") +
  xlab("hydrophone") +  ylab("Click count") +
  theme(axis.text.x=element_text(size=5),axis.title=element_text(size=9,face="bold")) +
  labs(fill='Direction') + theme(plot.title=element_text(hjust = 0.5))
```

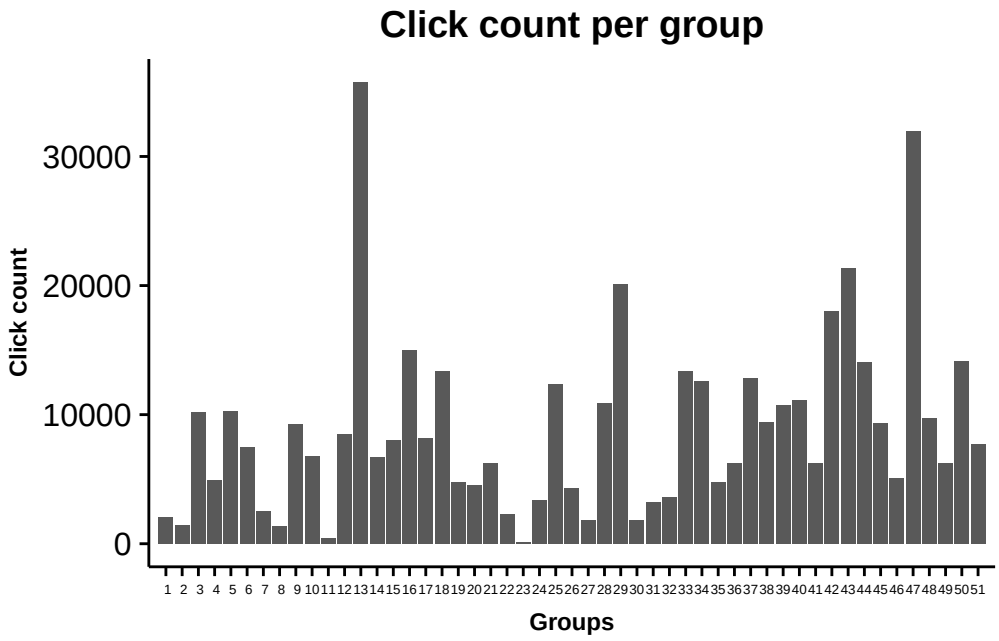



```
#Percentage of clicks detected by Whiskey hydrophones
pwh <- round(sum(gdatac$count[gdatac$whiskey==1])/sum(gdatac$count),2)*100

#Percentage of clicks detected by non-Whiskey hydrophones
pnwh <- round(sum(gdatac$count[gdatac$whiskey==0])/sum(gdatac$count),2)*100
```

Simply by looking at the picture above it is noticeable the 14 Whiskey hydrophones (out of 93 total hydrophones) account for a considerable amount of clicks, which happens to be approximately 34%, while non-Whiskey hydrophones account for 66%.

Bellow there is the number of clicks detected for each group.

```
ggplot(gdatac, aes(x = gID, y = count,colour)) +
  geom_bar(stat="identity") + scale_x_discrete(limits=c(1:51)) + labs(fill='Whiskey hyd') +
  ggtitle ("Click count per group") + xlab("Groups") +  ylab ("Click count") +
  theme(axis.text.x=element_text(size=5), axis.title=element_text(size=9,face="bold")) +
  theme(plot.title = element_text(hjust = 0.5)) + theme(aspect.ratio=3/5)
```

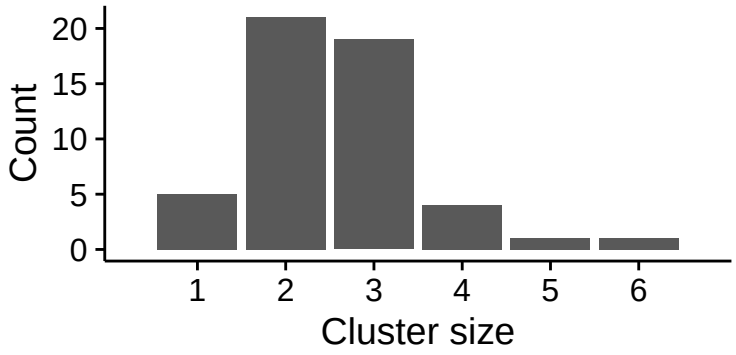



```
gn <- with(gdatac,tapply(X = count, INDEX = gID, FUN = sum))
gndata <-  order(gn, decreasing = TRUE)
```

The three groups with the most click counts are the numbers 13, 47, 43.

The figure bellow illustrates the response variable (Group/Cluster size) count.

```
ggplot(d4reg, aes(x = cs), geom = "Count")  + stat_count() +
  scale_x_discrete(name ="Cluster size", limits=c("1","2","3","4","5","6")) +
  ylab ("Count")
```



It is also important to highlight it is taken into account in this study whether or not a hydrophone is located on the edge. Since these type of hydrophones have a higher chance of capturing echolocations out of the considered area of 1291km^2, it is more likely that they incorporate false positive detections. Edge hydrophones include numbers 1, 2, 3, 15, 16, 17, 20, 24, 25, 30, 34, 35, 41, 42, 46, 53, 56, 61, 64, 69, 72, 77, 78, 80, 85, 88, 91, 92 and 93.

The response variable resembles a Poisson response. Although, if there are any signs of over dispersion it means a Negative Binomial approach may be useful. The mean count for the response variable is 2.57, while the variance is 0.97. Since the variance is smaller than the mean, there appears to be no over dispersion, but under dispersion instead. Additionally, the Negative Binomial approach raised warning messages as its

parameter was not able to converge and, therefore, will not be considered in the analysis. With this in mind, the Poisson distribution appears to reasonably fit the response variable.

For the sake of academic curiosity, and since the standard R functions (in particular the Chi-square test) do not support zero-truncated methods, we used an *ad hoc* approach to test if the Poisson response fits the modified response variable (cs0), which was used for non-truncated GLM.

```
#Defining sample size and the count for each response (using cs0)
nd <- length(d4reg$cs0)
a <- length(which(d4reg$cs0==0))
b <- length(which(d4reg$cs0==1))
c <- length(which(d4reg$cs0==2))
d <- length(which(d4reg$cs0==3))
e <- length(which(d4reg$cs0==4))
f <- length(which(d4reg$cs0==5))

#Creating a vector with the observed responses
x  <-  rep(0:5, times=c(a, b, c, d, e, f))

#Empirical Poisson response with the observed responses mean value
probs <- dpois(0:5, lambda=mean(x))
probp <- as.integer(probs*nd)
probp <- as.data.frame(probp)

#Creating another vector with the empirical responses
x1 <- c(rep(0,probp[1,]),rep(1,probp[2,]),rep(2,probp[3,]),
     rep(3,probp[4,]),rep(4,probp[5,]),rep(5,probp[6,]))
probp=as.data.frame(x1)

#Contabilizing the remaining probabilities in order to include them in the test
comp <- 1-sum(probs)

#Chi-square test with 10000 p-value iterations (the zero includes "comp")
chisq.test(x = c(a,b,c,d,e,f,0), p = c(probs, comp), simulate.p.value = TRUE, B = 10000)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value
##  (based on 10000 replicates)
##
## data:  c(a, b, c, d, e, f, 0)
## X-squared = 9.3333, df = NA, p-value = 0.154
```

It appears that the Poisson distribution fits well to the data.

2.3 Univariate analysis

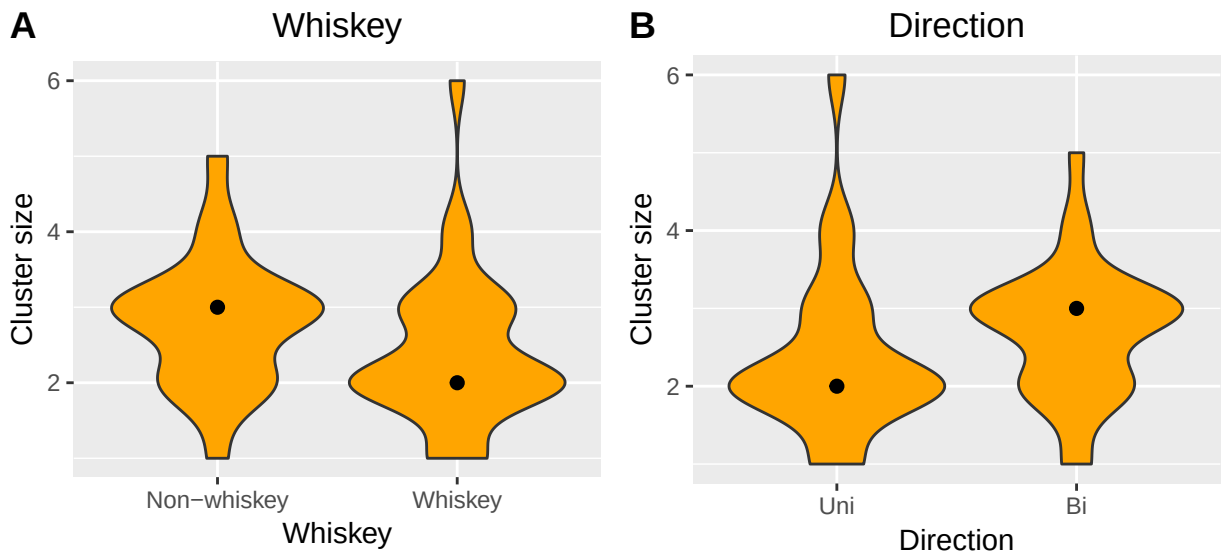
Before engaging in the modelling process it may be important to visualize how each explanatory variable behaves.

```
#Binary variables
uni1 <- ggplot(d4reg,aes(x = wisk, y = cs)) + geom_violin(fill = "orange") +
  stat_summary(fun.y=median, geom="point", size=2, color="black") + xlab("Whiskey") +
  ylab ("Cluster size") + scale_x_discrete(labels=c("Non-whiskey","Whiskey")) +
  theme_gray() + theme(plot.title = element_text(hjust = 0.5)) + ggtitle("Whiskey")
```

```
uni2 <- ggplot(d4reg,aes(x = direction, y = cs)) + geom_violin(fill = "orange") +
  stat_summary(fun.y=median, geom="point", size=2, color="black") + xlab("Direction") +
  ylab ("Cluster size") + scale_x_discrete(labels=c("Uni","Bi")) + theme_gray() +
  theme(plot.title = element_text(hjust = 0.5)) + ggtitle("Direction")

#Non-binary variables
uni3 <- ggplot(d4reg, aes(y = cs, x = meancount)) + geom_point(size=1) +
  ggtitle ("Click mean count") + scale_colour_manual(values=c("red", "blue")) +
   xlab ("Click meancount") + theme(plot.title = element_text(hjust = 0.5)) +
  theme_gray() + stat_smooth(method="glm",size=1) + ylab("Cluster size")

uni4 <- ggplot(d4reg, aes(y = cs, x = nhyd)) + geom_point(size=1) +
  ggtitle ("Click mean count") + scale_colour_manual(values=c("red", "blue")) +
   xlab ("Click meancount") + theme(plot.title = element_text(hjust = 0.5)) +
  theme_gray() + stat_smooth(method="glm",size=1) + ylab("Cluster size")

uni5 <- ggplot(d4reg, aes(y = cs, x = cdur)) + geom_point(size=1) +
  ggtitle ("Click mean count") + scale_colour_manual(values=c("red", "blue")) +
   xlab ("Click meancount") + theme(plot.title = element_text(hjust = 0.5)) +
  theme_gray() + stat_smooth(method="glm",size=1) + ylab("Cluster size")

uni6 <- ggplot(d4reg, aes(y = cs, x = nclicks)) + geom_point(size=1) +
  ggtitle ("Click mean count") + scale_colour_manual(values=c("red", "blue")) +
   xlab ("Click meancount") + theme(plot.title = element_text(hjust = 0.5)) +
  theme_gray() + stat_smooth(method="glm",size=1) + ylab("Cluster size")

uni7 <- ggplot(d4reg, aes(y = cs, x = crate)) + geom_point(size=1) +
  ggtitle ("Click mean count") + scale_colour_manual(values=c("red", "blue")) +
   xlab ("Click meancount") + theme(plot.title = element_text(hjust = 0.5)) +
  theme_gray() + stat_smooth(method="glm",size=1) + ylab("Cluster size")
```

The plots bellow illustrate each explanatory variable behaviour against the response variable.

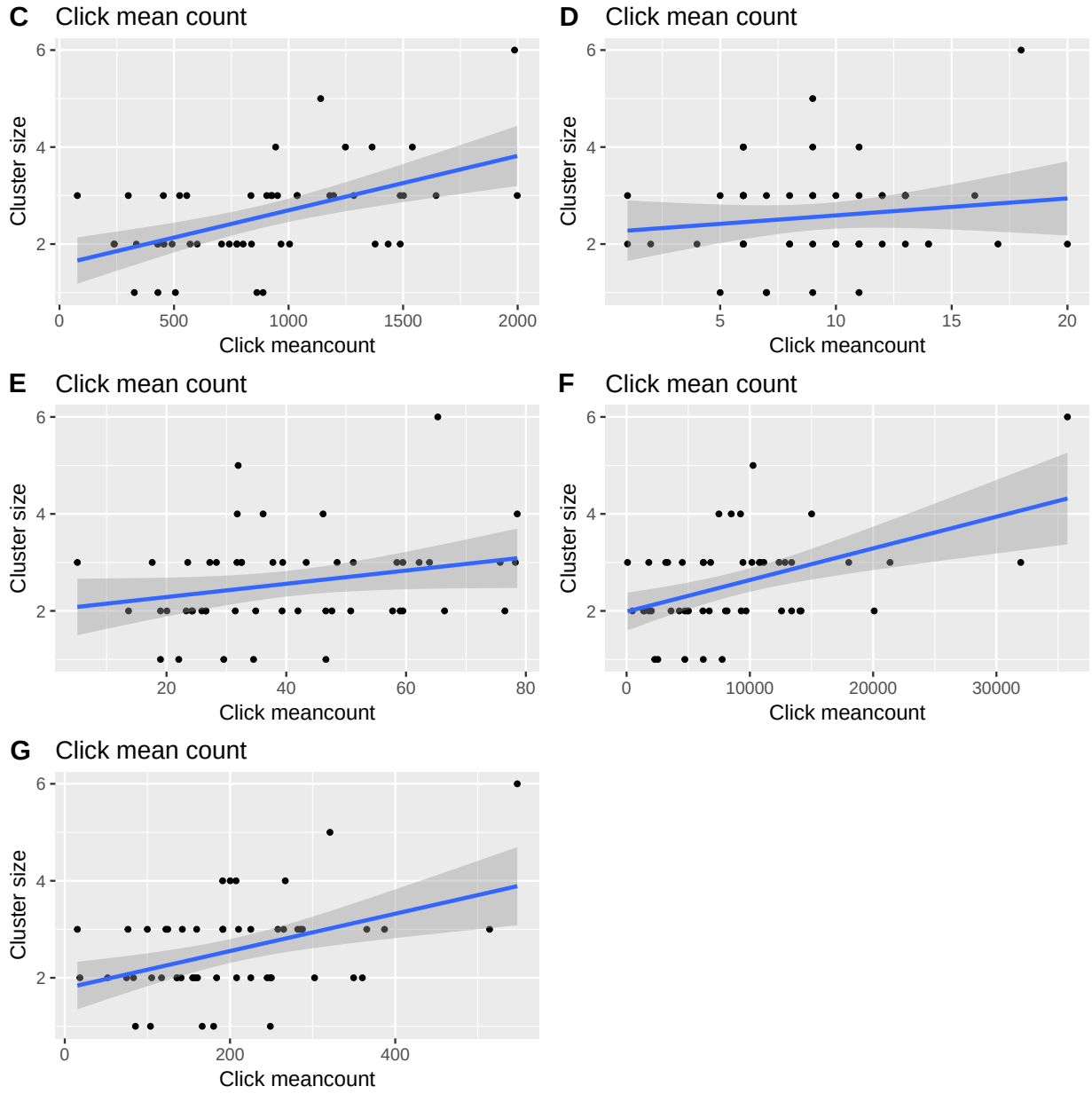
```
plot_grid(uni1, uni2,  labels = c("A", "B"))
```



It is noticeable that while the group size decreases from non-Whiskey to Whiskey hydrophones, it increases

from Uni to Bi-directional hydrophones. It may be important to highlight that all the Whiskey hydrophones are Uni-directional.

```
plot_grid(uni3, uni4, uni5, uni6,uni7, labels = c("C", "D", "E", "F", "G"),
          ncol = 2, nrow = 3)
```



All the other variables appear to increase along with the group size, specially **click mean count**, **number of clicks**, and **click rate**.

2.4 Interaction

It is important to analyse the relationship between the explanatory variables and the response variable. Interaction occurs when the effect of a explanatory variable on another one is not constant as the effect is not equal for different values the variable takes. This means a variable depends on the relationship between

the interacting variables and not the variables themselves; which may have significant implications for the interpretation of statistical models.

```
#Between nhyd and wisk
summary(lm(d4reg$cs ~ d4reg$nhyd * d4reg$wisk)) #not significant

#Between nhyd and direction
summary(lm(d4reg$cs ~ d4reg$nhyd * d4reg$direction)) #not significant

#Between nhyd and crate
summary(lm(d4reg$cs ~ d4reg$nhyd * d4reg$crate)) #not significant

#Between nhyd and cdur
summary(lm(d4reg$cs ~ d4reg$nhyd * d4reg$cdur)) #not significant

#Between nhyd and meancount
summary(lm(d4reg$cs ~ d4reg$nhyd * d4reg$meancount)) #possibly not significant (0.089)

#Between nhyd and nclicks
summary(lm(d4reg$cs ~ d4reg$nhyd * d4reg$nclicks)) #not significant

#--

#Between crate and wisk
summary(lm(d4reg$cs ~ d4reg$wisk * d4reg$crate)) #not significant

#Between crate and direction
summary(lm(d4reg$cs ~ d4reg$crate * d4reg$direction)) #not significant

#Between crate and cdur
summary(lm(d4reg$cs ~ d4reg$crate * d4reg$cdur)) #not significant

#Between crate and meancount
summary(lm(d4reg$cs ~ d4reg$crate * d4reg$meancount)) #not significant

#Between crate and nclicks
summary(lm(d4reg$cs ~ d4reg$crate * d4reg$nclicks)) #not significant

#--

#Between cdur and wisk
summary(lm(d4reg$cs ~ d4reg$wisk * d4reg$crate)) #not significant

#Between cdur and direction
summary(lm(d4reg$cs ~ d4reg$crate * d4reg$direction)) #not significant

#Between cdur and meancount
summary(lm(d4reg$cs ~ d4reg$cdur * d4reg$meancount)) #not significant

#Between cdur and nclicks
summary(lm(d4reg$cs ~ d4reg$cdur * d4reg$nclicks)) #not significant

#--
```

```r
#Between meancount and wisk
summary(lm(d4reg$cs ~ d4reg$meancount * d4reg$wisk)) #not significant

#Between meancount and direction
summary(lm(d4reg$cs ~ d4reg$meancount * d4reg$direction)) #not significant

#Between meancount and nclicks
summary(lm(d4reg$cs ~ d4reg$meancount * d4reg$nclicks)) #not significant

#--

#Between direction and wisk
summary(lm(cs ~ direction+ wisk+ direction * wisk, data=d4reg)) #not significant
```

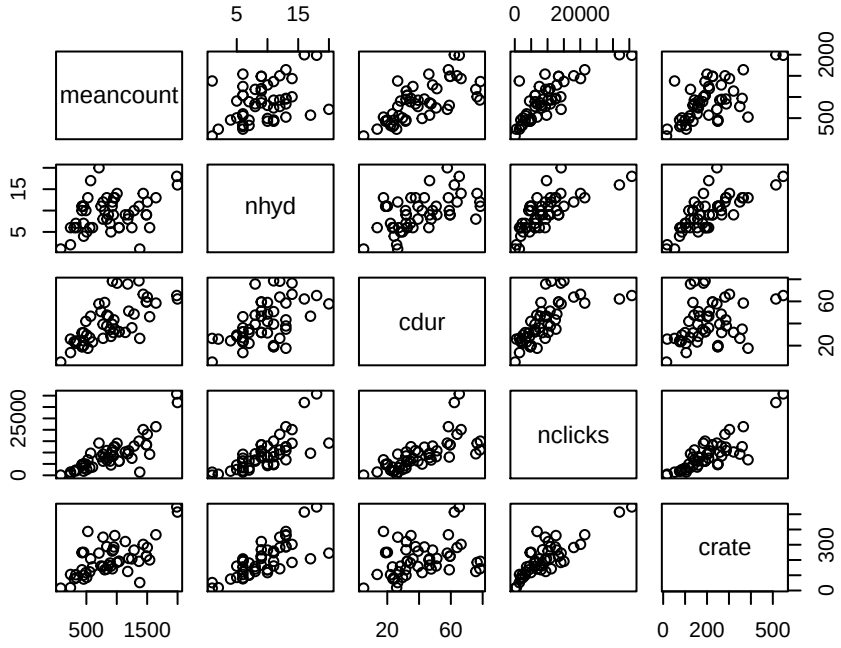
All the interactions are not significant for alpha=0.05. The only single significant interaction for alpha=0.01 is between **number of hydrophones** and **click mean count**.

2.5 Correlation

Correlation, whether causal or not, may indicate a predictive relationship that can be beneficial, since it may be possible to predict a variable from another one.

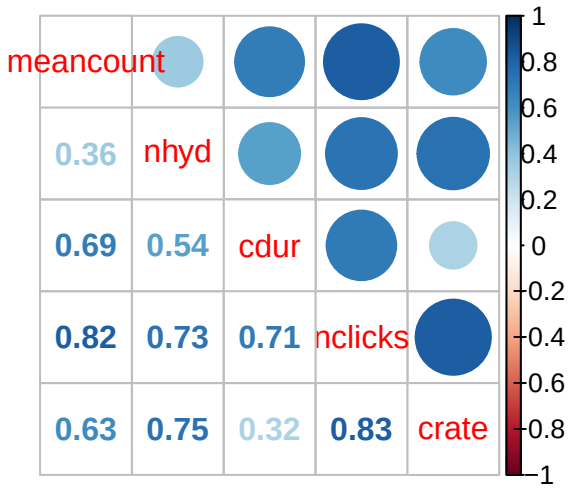
The figure below starts with illustrating each non-binary variable behaviour against the other.

```r
plot(d4reg[5:9])
```



The two plots bellow also regard the non-binary variables: the first one holds the estimated Pearson's coefficient values, and the second plot holds estimated Spearman's coefficient values.

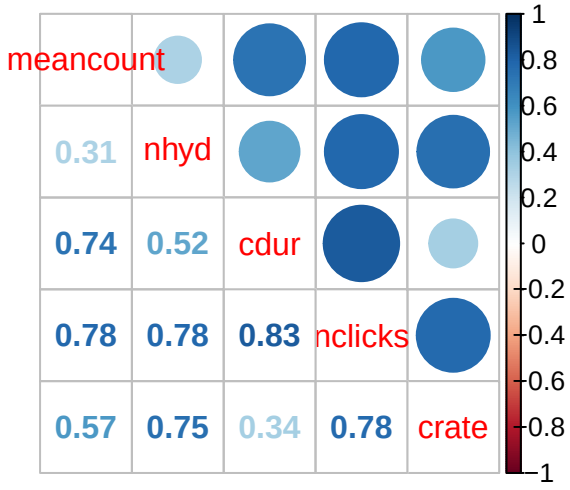
```
xcorpearson <- cor(d4reg[5:9], method = "pearson")
corrplot.mixed(xcorpearson)
```




```
#Estimated Pearson's coefficients
round(xcorpearson,2)
```

```
##           meancount nhyd cdur nclicks crate
## meancount      1.00 0.36 0.69    0.82  0.63
## nhyd           0.36 1.00 0.54    0.73  0.75
## cdur           0.69 0.54 1.00    0.71  0.32
## nclicks        0.82 0.73 0.71    1.00  0.83
## crate          0.63 0.75 0.32    0.83  1.00
```

```
xcorspear <- cor(d4reg[5:9], method = "spearman")
corrplot.mixed(xcorspear)
```



```r
#Estimated Spearman's coefficients
round(xcorspear,2)
```

```
##           meancount nhyd cdur nclicks crate
## meancount      1.00 0.31 0.74    0.78  0.57
## nhyd           0.31 1.00 0.52    0.78  0.75
## cdur           0.74 0.52 1.00    0.83  0.34
## nclicks        0.78 0.78 0.83    1.00  0.78
## crate          0.57 0.75 0.34    0.78  1.00
```

Bellow there are the Point-Biserial coefficients between the binary and non-binary variables.

```r
#nclicks & direction/wisk
ct1 <- biserial.cor(d4reg$nclicks,d4reg$direction,level=2)
ct2 <- biserial.cor(d4reg$nclicks,d4reg$wisk,level=2)

#cdur & direction/wisk
ct3 <- biserial.cor(d4reg$cdur,d4reg$direction,level=2)
ct4 <- biserial.cor(d4reg$cdur,d4reg$wisk,level=2)

#crate & direction/wisk
ct5 <- biserial.cor(d4reg$crate,d4reg$direction,level=2)
ct6 <- biserial.cor(d4reg$crate,d4reg$wisk,level=2)

#nhyd & direction/wisk
ct7 <- biserial.cor(d4reg$nhyd,d4reg$direction,level=2)
ct8 <- biserial.cor(d4reg$nhyd,d4reg$wisk,level=2)

#meancount & direction/wisk
ct9 <- biserial.cor(d4reg$meancount,d4reg$direction,level=2)
ct10 <- biserial.cor(d4reg$meancount,d4reg$wisk,level=2)

cttable <- round(matrix(c(ct1,ct2,ct3,ct4,ct5,ct6,ct7,ct8,ct9,ct10),ncol=2,byrow=TRUE),2)
colnames(cttable) <- c("direction","wisk")
rownames(cttable) <- c("nclicks","cdur","crate","nhyd","meancount")
cttable
```

```
##           direction  wisk
## nclicks       -0.16  0.20
## cdur          -0.04  0.05
## crate         -0.18  0.31
## nhyd          -0.36  0.48
## meancount      0.11 -0.08
```

And there is also the Phi coefficient between both binary variables.

```r
#Phi coefficient
bitab <- with(d4reg, table(direction, wisk))
bvar <- matrix(c(phi(bitab)))
colnames(bvar) <- c("direction")
rownames(bvar) <- c("wisk")
bvar
```

```
##      direction
## wisk     -0.68
```

All the non-binary variables appear to be positively correlated. The two duos whose coefficients are closer to

zero are **meancount & nhyd**, and **cdur & crate**.

Additionally, some non-binary variables appear to correlate with **direction** and **wisk**, specially **nhyd** (positive correlation with **wisk**, and negative correlation with **direction**).

Finally, both binary variables **wisk** and **direction** appear to be negatively correlated.

Despite the correlation between variables, let us proceed with the modelling to see which of them appear to be statistically significant.

2.6 Modelling

In order to evaluate potential differences between the groups with different confidence levels, a new dataset only containing the 43 groups with a confidence level of 1 was created.

Furthermore, when attempting to fit a Negative Binomial response, a warning message prompts stating convergence was not obtained for the *theta* value. According to literature, for very large *theta* values the coefficient estimates are close to a Poisson distribution. (Klugman *et al*, 2004).

```
#Creating a dataset with only the groups with conf=1
d4regnew <- d4reg[- c(which(d4reg$conf!=1)),]
```

2.6.1 Non-truncated approach

We start with the non-truncated GLM method.

Considering all the 51 observations:

```
model1 <- glm(cs0~meancount+cdur+nhyd+nclicks+crate+factor(wisk)
           +factor(direction), family=poisson, data=d4reg)
drop1(model1,test="F") #remove direction

model2 <- glm(cs0~meancount+cdur+nhyd+nclicks+crate+factor(wisk), family=poisson,
              data=d4reg)
drop1(model2,test="F") #remove nclicks

model3 <- glm(cs0~meancount+cdur+nhyd+crate+factor(wisk), family=poisson, data=d4reg)
drop1(model3,test="F") #remove meancount

model4 <- glm(cs0~cdur+nhyd+crate+factor(wisk),family=poisson, data=d4reg)
drop1(model4,test="F") #remove wisk

model5 <- glm(cs0~nhyd+crate+cdur,family=poisson,data=d4reg)
drop1(model5,test="F") #remove cdur

model6 <- glm(cs0~crate+nhyd,family=poisson,data=d4reg)
drop1(model6,test="F") #remove nhyd

model7 <- glm(cs0~crate,family=poisson,data=d4reg)
```

- model1: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.
- model2: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey.
- model3: mean count + click duration + number of hydrophones + click rate + whiskey.

- model4: click duration + number of hydrophones + click rate + whiskey.
- model5: click duration + number of hydrophones + click rate.
- model6: number of hydrophones + click rate.
- model7: click rate.

```
aicglm1 <- c(AIC(model1), AIC(model2), AIC(model3), AIC(model4), AIC(model5),
          AIC(model6), AIC(model7))
glm1 <- c("model1","model2","model3","model4","model5","model6","model7")
glmAIC1 <- matrix(aicglm1,ncol=1,byrow=TRUE)
rownames(glmAIC1) <- glm1
colnames(glmAIC1) <- c("AIC")
kable(glmAIC1, caption = "AIC values for the non-truncated GLM, n=51")
```

Now, non-truncated GLM with only the 43 groups with conf=1:

```
mnew1 <- glm(cs0~meancount+cdur+nhyd+nclicks+crate+factor(wisk)
          +factor(direction), family=poisson, data=d4regnew)
drop1(mnew1,test="F") #remove meancount

mnew2 <- glm(cs0~cdur+nhyd+nclicks+crate+factor(wisk)+factor(direction), family=poisson,
          data=d4regnew)
drop1(mnew2,test="F") #remove direction

mnew3 <- glm(cs0~cdur+nhyd+nclicks+crate+factor(wisk), family=poisson, data=d4regnew)
drop1(mnew3,test="F") #remove nclicks

mnew4 <- glm(cs0~cdur+nhyd+crate+factor(wisk),family=poisson, data=d4regnew)
drop1(mnew4,test="F") #remove nhyd

mnew5 <- glm(cs0~cdur+crate+factor(wisk),family=poisson,data=d4regnew)
drop1(mnew5,test="F") #remove cdur

mnew6 <- glm(cs0~crate+factor(wisk),family=poisson,data=d4regnew)
drop1(mnew6,test="F") #remove wisk

mnew7 <- glm(cs0~crate,family=poisson,data=d4regnew)
```

- mnew1: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.
- mnew2: click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.
- mnew3: click duration + number of hydrophones + number of clicks + click rate + whiskey.
- mnew4: click duration + number of hydrophones + click rate + whiskey.
- mnew5: click duration + click rate + whiskey.
- mnew6: click rate + whiskey.
- mnew7: click rate.

```
aicglm2 <- c(AIC(mnew1), AIC(mnew2), AIC(mnew3), AIC(mnew4), AIC(mnew5),
          AIC(mnew6), AIC(mnew7))
glm2 <- c("mnew1","mnew2","mnew3","mnew4","mnew5","mnew6","mnew7")
glmAIC2 <- matrix(aicglm2,ncol=1,byrow=TRUE)
rownames(glmAIC2) <- glm2
```

```r
colnames(glmAIC2) <- c("AIC")
kable(glmAIC2, caption = "AIC values for the non-truncated GLM, n=43")
```

Table 1: AIC values for the non-truncated GLM, n=43

	AIC
mnew1	128.9031
mnew2	126.9032
mnew3	124.9752
mnew4	123.1531
mnew5	122.0444
mnew6	120.4265
mnew7	122.1139

When considering all the 51 groups, the 3 best models include **cdur**, **crate**, and/or **nhyd**. Meanwhile, when using the dataset with only 43 groups the variable **nhyd** is replaced by **wisk**. Although, they both share a model only containing the variable **crate**.

2.6.2 Zero-truncated approach

In this section, zero-truncated GLM will be employed.

Bellow there are the results considering all the 51 observations:

```r
mp1 <- vglm(cs ~ meancount + cdur + nhyd + nclicks + crate + factor(wisk)
            + factor(direction), family = pospoisson, data = d4reg)
summary(mp1)

mp2 <- update(mp1, . ~ . - factor(direction))
summary(mp2)

mp3 <- update(mp2, . ~ . - nclicks)
summary(mp3)

mp4 <- update(mp3, . ~ . - meancount)
summary(mp4)

mp5 <- update(mp4, . ~ . - factor(wisk))
summary(mp5)

mp6 <- update(mp5, . ~ . - cdur)
summary(mp6)

mp7 <- update(mp6, . ~ . - nhyd)
summary(mp7)
```

- mp1: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.
- mp2: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey.
- mp3: mean count + click duration + number of hydrophones + click rate + whiskey.
- mp4: click duration + number of hydrophones + click rate + whiskey.

- mp5: click duration + number of hydrophones + click rate.
- mp6: number of hydrophones + click rate.
- mp7: click rate.

```
aict1 <- c(AIC(mp1), AIC(mp2), AIC(mp3), AIC(mp4), AIC(mp5), AIC(mp6), AIC(mp7))
tglm1 <- c("mp1","mp2","mp3","mp4","mp5","mp6","mp7")
mtAIC1 <- matrix(aict1,ncol=1,byrow=TRUE)
rownames(mtAIC1) <- tglm1
colnames(mtAIC1) <- c("AIC")

kable(mtAIC1, caption = "AIC values for the zero-truncated GLM,  n=51")
```

Table 2: AIC values for the zero-truncated GLM, n=51

	AIC
mp1	157.5694
mp2	155.5757
mp3	153.5909
mp4	151.6625
mp5	150.4935
mp6	150.7108
mp7	150.5499

And now, only with the 43 groups:

```
mpn1 <- vglm(cs ~ meancount + cdur + nhyd + nclicks + crate + factor(wisk)
             + factor(direction), family = pospoisson, data = d4regnew)
summary(mpn1)

mpn2 <- update(mpn1, . ~ . - meancount)
summary(mpn2)

mpn3 <- update(mpn2, . ~ . - factor(direction))
summary(mpn3)

mpn4 <- update(mpn3, . ~ . - nclicks)
summary(mpn4)

mpn5 <- update(mpn4, . ~ . - nhyd)
summary(mpn5)

mpn6 <- update(mpn5, . ~ . - cdur)
summary(mpn6)

mpn7 <- update(mpn6, . ~ . - factor(wisk))
summary(mpn7)
```

- mpn1: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.
- mpn2: click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.
- mpn3: click duration + number of hydrophones + number of clicks + click rate + whiskey.

- mpn4: click duration + number of hydrophones + click rate + whiskey.
- mpn5: click duration + click rate + whiskey.
- mpn6: click rate + whiskey.
- mpn7: click rate.

```
aict2 <- c(AIC(mp1), AIC(mp2), AIC(mp3), AIC(mp4), AIC(mp5), AIC(mp6), AIC(mp7))
tglm2 <- c("mp1","mp2","mp3","mp4","mp5","mp6","mp7")
mtAIC2 <- matrix(aict2,ncol=1,byrow=TRUE)
rownames(mtAIC2) <- tglm2
colnames(mtAIC2) <- c("AIC")

kable(mtAIC2, caption = "AIC values for the zero-truncated GLM,  n=43")
```

Table 3: AIC values for the zero-truncated GLM, n=43

	AIC
mp1	157.5694
mp2	155.5757
mp3	153.5909
mp4	151.6625
mp5	150.4935
mp6	150.7108
mp7	150.5499

Note that both non-truncated and zero-truncated GLM select the same variables when considering the same number of observations.

The model with the single variable **crate** appears to be the best one to describe the response variable.

2.6.2.1 Analysing the model

2.6.2.1.1 Residuals

First, we test for the normality of the residuals.

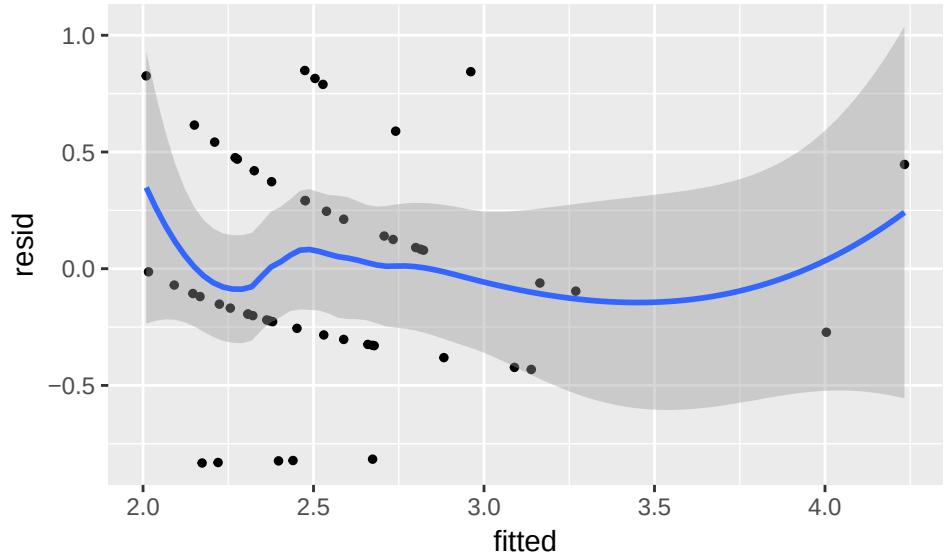
```
mp7 <- vglm(cs ~ crate, family = pospoisson, data = d4reg)
shapiro.test(residuals(mp7,type="pearson"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mp7, type = "pearson")
## W = 0.97532, p-value = 0.3626
```

Normality is not rejected.

The figure bellow illustrates the residuals *vs* fitted values behaviour, with a scatter plot smoother in grey:

```
theme_set(theme_grey())
output <- data.frame(resid = resid(mp7), fitted = fitted(mp7))
ggplot(output, aes(fitted, resid)) +
  geom_jitter(size=1)+ stat_smooth(method="loess")
```



Checking if the residuals and the fitted values are correlated:

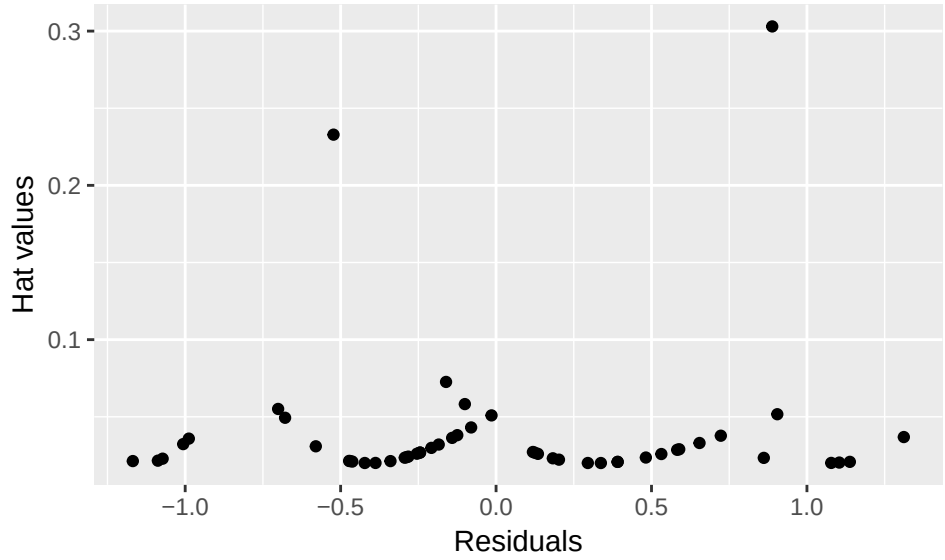
```
cor.test(fitted(mp7), resid(mp7))
```

```
##
##  Pearson's product-moment correlation
##
## data:  fitted(mp7) and resid(mp7)
## t = -0.1068, df = 49, p-value = 0.9154
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2896211  0.2614278
## sample estimates:
##         cor
## -0.01525496
```

They appear to not be correlated.

2.6.2.1.2 Hat Values

```
output2 <- data.frame(resid = resid(mp7, "pearson"), hatv = hatvalues(mp7))
names(output2)[2] <-"hatv"
ggplot(output2, aes(resid, hatv)) + geom_point() + ylab("Hat values") +  xlab ("Residuals")
```




```
#Obtaining the highest hat values
nlength <- length(d4reg$gID)
hvthres <- (2*sum(hatvalues(mp7)))/nlength
hv <- which(output2$hatv>hvthres)
```

The observations with the highest hat values are 13, 47.

In order to see how much influence these points have, let us rebuild the model without them.

```
#Building a new dataset without the influential values
d4regnewh1=d4reg[- c(hv),]

mpnh1 <- vglm(cs ~ meancount + cdur + nhyd + nclicks + crate + factor(wisk)
              + factor(direction), family = pospoisson, data = d4regnewh1)
summary(mpnh1)

mpnh2 <- update(mpnh1, . ~ . - factor(direction))
summary(mpnh2)

mpnh3 <- update(mpnh2, . ~ . - nclicks)
summary(mpnh3)

mpnh4 <- update(mpnh3, . ~ . - meancount)
summary(mpnh4)

mpnh5 <- update(mpnh4, . ~ . - factor(wisk))
summary(mpnh5)

mpnh6 <- update(mpnh5, . ~ . - cdur)
summary(mpnh6)

mpnh7 <- update(mpnh6, . ~ . - nhyd)
summary(mpnh7)
```

- mpnh1: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.

- mpnh2: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey.
- mpnh3: mean count + click duration + number of hydrophones + click rate + whiskey.
- mpnh4: click duration + number of hydrophones + click rate + whiskey.
- mpnh5: click duration + number of hydrophones + click rate.
- mpnh6: number of hydrophones + click rate.
- mpnh7: click rate.

```
aich1 <- c(AIC(mpnh1), AIC(mpnh2), AIC(mpnh3), AIC(mpnh4), AIC(mpnh5), AIC(mpnh6),
           AIC(mpnh7))
mhat1 <- c("mpnh1","mpnh2","mpnh3","mpnh4","mpnh5","mpnh6","mpnh7")
mhAIC1 <- matrix(aich1,ncol=1,byrow=TRUE)
rownames(mhAIC1) <- mhat1
colnames(mhAIC1) <- c("AIC")

kable(mhAIC1, caption = "AIC values for zero-truncated GLM,  n=49")
```

Table 4: AIC values for zero-truncated GLM, n=49

	AIC
mpnh1	149.9548
mpnh2	147.9566
mpnh3	145.9748
mpnh4	144.0410
mpnh5	142.7658
mpnh6	142.9766
mpnh7	142.9038

Now, modelling without the hat values and only with the observations with conf=1:

```
hv2 <- which(d4reg$conf!=1 | hatvalues(mp7)>(2*sum(hatvalues(mp7))/51))
d4regnewh2=d4reg[- hv2,]

mpnhc1 <- vglm(cs ~ meancount + cdur + nhyd + nclicks + crate + factor(wisk) +
                 factor(direction), family =pospoisson, data = d4regnewh2)
summary(mpnhc1)

mpnhc2 <- update(mpnhc1, . ~ . - factor(direction))
summary(mpnhc2)

mpnhc3 <- update(mpnhc2, . ~ . - nhyd)
summary(mpnhc3)

mpnhc4 <- update(mpnhc3, . ~ . - cdur)
summary(mpnhc4)

mpnhc5 <- update(mpnhc4, . ~ . - nclicks)
summary(mpnhc5)

mpnhc6 <- update(mpnhc5, . ~ . - meancount)
summary(mpnhc6)
```

```
mpnhc7 <- update(mpnhc6, . ~ . - crate)
summary(mpnhc7)

mpnhc8 <- update(mpnhc7, . ~ . - factor(wisk))
summary(mpnhc8)
```

- mpnhc1: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey + direction.
- mpnhc2: mean count + click duration + number of hydrophones + number of clicks + click rate + whiskey.
- mpnhc3: mean count + click duration + number of clicks + click rate + whiskey.
- mpnhc4: mean count + number of clicks + click rate + whiskey.
- mpnhc5: mean count + click rate + whiskey.
- mpnhc6: click rate + whiskey.
- mpnhc7: click rate.

```
aich2 <- c(AIC(mpnhc1), AIC(mpnhc2), AIC(mpnhc3), AIC(mpnhc4),
        AIC(mpnhc5), AIC(mpnhc6), AIC(mpnhc7))
mhat2 <- c("mpnhc1","mpnhc2","mpnhc3","mpnhc4","mpnhc5","mpnhc6","mpnhc7")
mhAIC2 <- matrix(aich2,ncol=1,byrow=TRUE)
rownames(mhAIC2) <- mhat2
colnames(mhAIC2) <- c("AIC")

kable(mhAIC2, caption = "AIC values for zero-truncated GLM,  n=43")
```

Table 5: AIC values for zero-truncated GLM, n=43

	AIC
mpnhc1	128.3319
mpnhc2	126.3888
mpnhc3	124.4536
mpnhc4	123.0695
mpnhc5	121.2859
mpnhc6	119.9851
mpnhc7	120.9000

We may see that removing more variables on an already small dataset leads to new variables to be chosen, and smaller coefficients of significance. Therefore, in order to include all the information, no observations were removed and the analysis proceeded with the model **mp7**.

3 The Density Estimation Dataset.

A second dataset (with unknown group sizes) will be used to estimate group sizes based on the model that related group size to acoustic footprint, which will then allow the density estimation over time. It contemplates three different time periods from 2011, covering a total of 113 days. However, 4 out of 113 these days were only partially sampled. Given our objective of producing density estimates per day, it is simpler to consider only the 109 days for which we have 24 hours of recording, and hence these incomplete days were discarded from further analysis.

The considered time periods are: (1) from the 28th of April to the 27th of June; (2) from the 20th of October to the 6th of November; and (3) from the 2nd to the 31st of December.

The code bellow describes how it was built:

```
dados <- read.table("dadostodos.txt", header=T)
dados.ghyd <- dados

#Sort the data by group number, then by hydrophone
dados.ghyd <- dados.ghyd[order(dados.ghyd$GroupNum,dados.ghyd$hyd),]
#---------------------------------------------------------
#Some hydrophones are included in multiple rows in the same group
#Put all in the same row. Creating an unique indicator
dados.ghyd$ghyd <- with(dados.ghyd,paste(GroupNum,hyd,sep="."))

#Now get the GroupNum
groupnum <- with(dados.ghyd,tapply(X=GroupNum, INDEX=ghyd, FUN = mean))

#Now get the hyd
hyd <- with(dados.ghyd,tapply(X=hyd, INDEX=ghyd, FUN = mean))

#Now get the total click count
clickcnt <- with(dados.ghyd,tapply(X=clickcnt, INDEX=ghyd, FUN = sum))

#Now get the start - i.e. the minimum
start <- with(dados.ghyd,tapply(X=start, INDEX=ghyd, FUN = min))

#Now get the end - i.e. the maximum
end <- with(dados.ghyd,tapply(X=end, INDEX=ghyd, FUN = max))

#Now get the minimum ici (this might be the best group size predictor)
mici <- with(dados.ghyd,tapply(X=ici, INDEX=ghyd, FUN = min))

#Now get how many times that hyd was repeated
hydreps <- with(dados.ghyd,tapply(X=hyd, INDEX=ghyd, FUN = length))

#Finally, bundle in the same data frame
dados.nrhyds <- data.frame(GroupNum=groupnum, hyd=hyd, clickcnt=clickcnt, start=start,
                           end=end,mici=mici)

#Lets create a new column, with the time duration of the click detection
dados.nrhyds["clickdur"] <- NA
dados.nrhyds$clickdur <- dados.nrhyds$end-dados.nrhyds$start

#And sort again as the tapply function messes up with sorting
dados.nrhyds <- dados.nrhyds[order(dados.nrhyds$GroupNum,dados.nrhyds$hyd),]

#Now add whether phones are edge phone or not
edgehyds <- c(1,2,3,15,16,17,20,24,25,30,34,35,41,42,46,53,56,61,64,69,
              72,77,78,80,85,88,91,92,93)
dados.nrhyds$edge <- dados.nrhyds$hyd %in% edgehyds

#Adding a whiskey/non-whiskey column
dados.nrhyds["whiskey"] <- dados.nrhyds$hyd
dados.nrhyds$whiskey <- car::recode(dados.nrhyds$whiskey, "1:14=TRUE")
```

```
dados.nrhyds$whiskey <- car::recode(dados.nrhyds$whiskey, "15:93=FALSE")
whiskey01 <- c(1)
dados.nrhyds$whiskey <- dados.nrhyds$whiskey %in% whiskey01

#Adding a bi/uni directional column
dados.nrhyds["direction"] <- dados.nrhyds$hyd
dados.nrhyds$direction <- car::recode(dados.nrhyds$direction, "1=0 ; 15=1; 20=1; 30=1;
41=1; 42=1; 45=1; 56=1; 58=1; 61=1; 69=1; 72=1; 75=1;
78=1; 88=1; 91=1; 93=1")
dados.nrhyds$direction <- car::recode(dados.nrhyds$direction, "2:93=0")
direction01 <- c(1)
dados.nrhyds$direction <- dados.nrhyds$direction %in% direction01

#Obtaining some relevant variables by group

#Getting the group's ID
groupnum2 <- with(dados.nrhyds,tapply(X=GroupNum, INDEX=GroupNum, FUN = mean))

#Number of hyds it was detected at
nhyd <- with(dados.nrhyds,tapply(X=hyd, INDEX=GroupNum, FUN = length))

#Total clicks
nclick <- with(dados.nrhyds,tapply(X=clickcnt, INDEX=GroupNum, FUN = sum))

#Get the start - i.e. the minimum
start <- with(dados.nrhyds,tapply(X=start, INDEX=GroupNum, FUN = min))

#Get the end - i.e. the maximum
end <- with(dados.nrhyds,tapply(X=end, INDEX=GroupNum, FUN = max))

#Now get the minimum ici
ici <- with(dados.nrhyds,tapply(X=mici, INDEX=GroupNum, FUN = min))

#And finally, define whether all hydrophones were edge phones
#which happens to be the case if the minimum value on the
#edge variable is 0 (if there's at least a non edge phone it becomes 0)
edge <- with(dados.nrhyds,tapply(X=edge, INDEX=GroupNum, FUN = min))

#Create an object to hold the data by groups
dados.groups <- data.frame(GroupNum=groupnum2,nhyd=nhyd,clickcnt=nclick,start=start,
                           end=end,ici=ici,edge=edge)

#Select groups which were detected on a single hydrophone only
#these are likely false positives
dados.groups$unihyd <- ifelse(dados.groups$nhyd==1,1,0)

#Select groups for which less than tresh clicks were detected
tresh <- 400
dados.groups$tresh <- ifelse(dados.groups$clickcnt<tresh,1,0)

#The following variable can be used to select
#only those thought to be true positives
dados.groups$tps <- with(dados.groups,edge+unihyd+tresh==0)
```

```
#Selecting that as a separate data frame
dados.filtered <- dados.groups[dados.groups$tps==1,]

#Time periods
dados.filtered["period"] <- dados.filtered$start
dados.filtered$period <- car::recode(dados.filtered$period, "1:15200=1")
dados.filtered$period <- car::recode(dados.filtered$period, "15201:15300=2")
dados.filtered$period <- car::recode(dados.filtered$period, "15301:15400=3")

#Click duration
dados.filtered["clickdur"] <- NA
dados.filtered$clickdur <- dados.filtered$end-dados.filtered$start
#It is in days, lets put it in minutes
dados.filtered$clickdur <- (dados.filtered$clickdur)*24 #it is now in hours
dados.filtered$clickdur <- (dados.filtered$clickdur)*60 #it is now in minutes

#Click rate
dados.filtered["crate"] <- NA
dados.filtered$crate <- (dados.filtered$clickcnt)/(dados.filtered$clickdur)

#Cluster size
dados.filtered["cs"] <- NA

#Date
dados.filtered["date"] <- NA
dados.filtered["date"] <- as.Date(dados.filtered$start, origin = "1970-01-01")

#Julian date
dados.filtered["day"] <- NA
tmp <- as.POSIXlt(dados.filtered$date, format = "%y%d%b")
tmp <- format(tmp, "%j")
dados.filtered$day <- tmp

#Date with hours
startd <- dados.filtered$start
tmd <- as.POSIXlt(startd*60*60*24, origin="1970-01-01", tz = "UTC")
dados.filtered$time <- tmd
dados.filtered$time[1]

#Removing days without 24 hours, we already know their numbers
ddf <- c(which(dados.filtered$day==117 | dados.filtered$day==292
            | dados.filtered$day==311 | dados.filtered$day==335))

dados.filtered <- dados.filtered[-c(ddf),]

period1 <- dados.filtered[dados.filtered$period==1,]
period2 <- dados.filtered[dados.filtered$period==2,]
period3 <- dados.filtered[dados.filtered$period==3,]
```

Now, using the previous model to predict the group size:

```
dados.filtered["cs"] <- predict(mp7, dados.filtered, type="response")
```

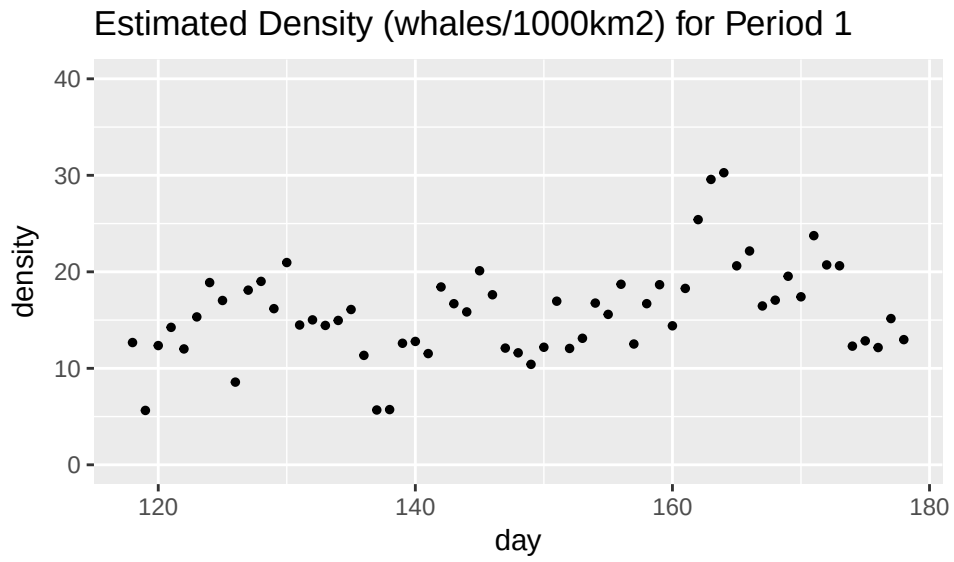
And creating a new dataset regarding the information per day:

```r
#Cluster size mean for each day
meancs <- with(dados.filtered,tapply(X=cs, INDEX=day, FUN = mean))


dados.filtered$time <- as.POSIXct(dados.filtered$time,format = "%d%m%Y %H:%M:%S",tz="UTC")
dados.filtered$time <- ymd_hms(dados.filtered$time)

as.numeric(difftime(dados.filtered$time[2], dados.filtered$time[1], tz="UTC",
         units = c("hours")))
t.str <- strptime(dados.filtered$time, "%Y-%m-%d %H:%M:%S", tz="UTC")
t.lub <- ymd_hms(dados.filtered$time)

#Extract decimal hours
h.str <- as.numeric(format(t.str, "%H")) +
  as.numeric(format(t.str, "%M"))/60

#Adding a collumn for decimal hours
dados.filtered$hours <- h.str

#The amount of hours per day the click detection occurred
#(just for curiosity, wont be used)
maxhours <- with(dados.filtered,tapply(X=hours, INDEX=day, FUN = max))
minhours <- with(dados.filtered,tapply(X=hours, INDEX=day, FUN = min))
inthours <- maxhours-minhours

#The number of groups detected per day
ngday <- as.numeric(with(dados.filtered,tapply(X=GroupNum, INDEX=day, FUN = length)))

#Creating new data frame uniquely for days
udngroup <- with(dados.filtered,tapply(X=GroupNum, INDEX=day, FUN = length))

#Number of groups for each day
udhyd <- with(dados.filtered,tapply(X=nhyd, INDEX=day, FUN = sum))
udnclick <- with(dados.filtered,tapply(X=clickcnt, INDEX=day, FUN = sum))
udici <- with(dados.filtered,tapply(X=ici, INDEX=day, FUN = min))
udperiod <- with(dados.filtered,tapply(X=period, INDEX=day, FUN = mean))
udcrate <- with(dados.filtered,tapply(X=crate, INDEX=day, FUN = mean))
udcs <- with(dados.filtered,tapply(X=cs, INDEX=day, FUN = mean))
uday <- unique(dados.filtered$day)

rd=0.36
area=1291
intdays <- inthours/24
Nh <- (ngday*meancs)/(rd*inthours) #abundance per hour
Nd <- (ngday*meancs)/(rd*intdays) #abundance per day

#Assembling the new dataset
dados.eachday <- data.frame(groups=as.numeric(udngroup), nhyd=as.numeric(udhyd),
                            nclick=as.numeric(udnclick), ici=as.numeric(udici),
                            period=as.numeric(udperiod), crate=as.numeric(udcrate),
                            mcs=as.numeric(udcs), day=as.numeric(uday),
                            time=as.numeric(minhours), etime=as.numeric(maxhours),
                            ttime=as.numeric(inthours))
```

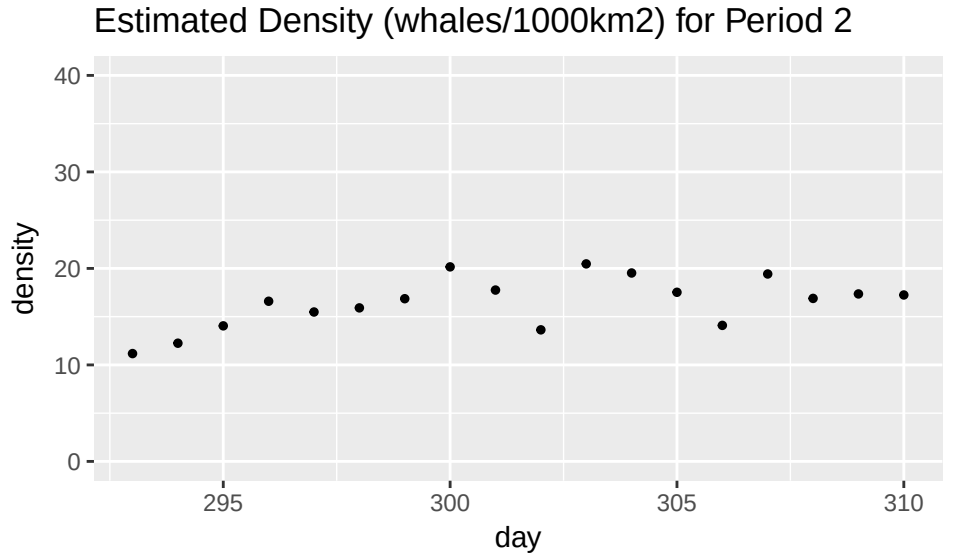
```
dados.eachday$abundancy <- ((dados.eachday$groups)*(dados.eachday$mcs))/(rd*24)
dados.eachday$density <- dados.eachday$abundancy/area*1000
```

Now, the plots for each time period illustrating the estimated density:

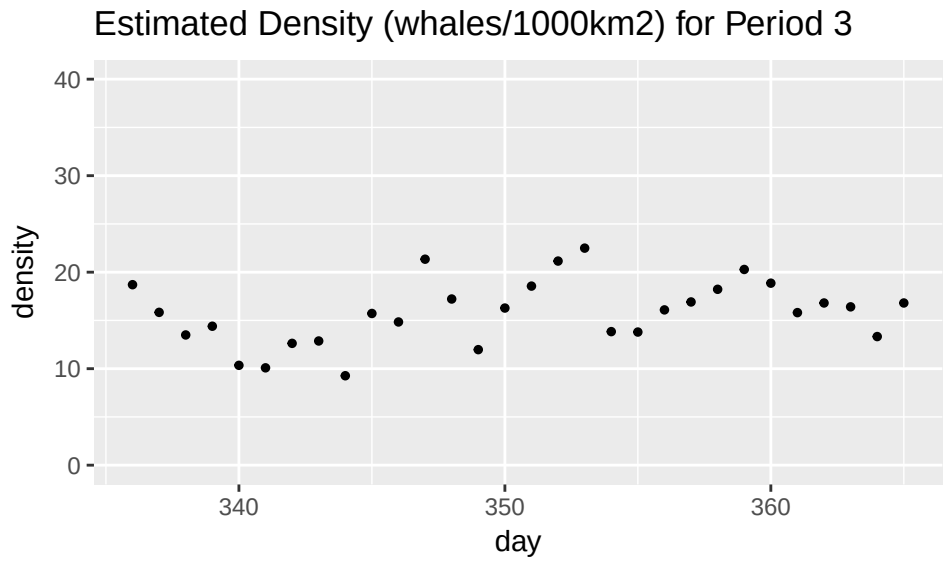
```
plot.ed1 <- ggplot(data=dados.eachday[dados.eachday$period==1,],aes(x= day, y = density))
plot.ed1 + geom_point(size=1) + scale_y_continuous(limits = c(0, 40)) +
  ggtitle("Estimated Density (whales/1000km2) for Period 1")
```




```
plot.ed2 <- ggplot(data=dados.eachday[dados.eachday$period==2,],aes(x= day, y = density))
plot.ed2 + geom_point(size=1) + scale_y_continuous(limits = c(0, 40)) +
  ggtitle("Estimated Density (whales/1000km2) for Period 2")
```




```
plot.ed3 <- ggplot(data=dados.eachday[dados.eachday$period==3,],aes(x= day, y = density))
plot.ed3 + geom_point(size=1) + scale_y_continuous(limits = c(0, 40)) +
  ggtitle("Estimated Density (whales/1000km2) for Period 3")
```



4 Bootstrap

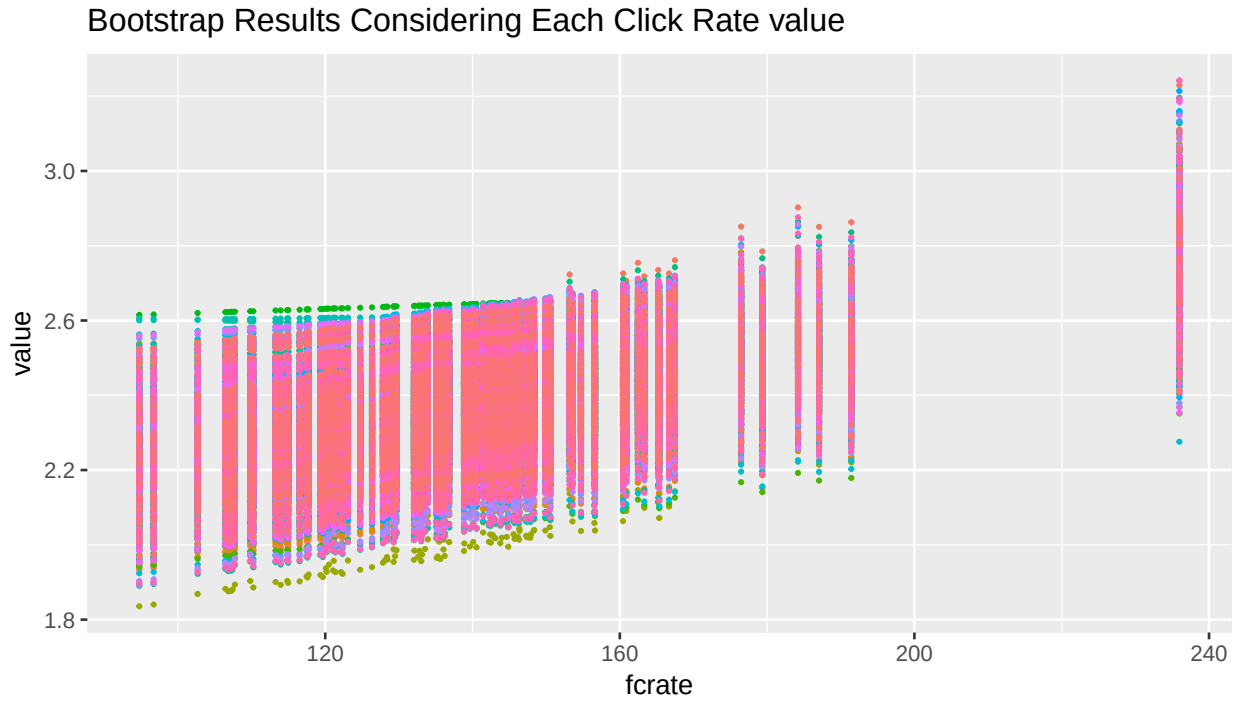
The final task is to propagate the variance in the model of group size thorough the estimates of variance of density per day. Therefore, the modelling dataset will be re-sampled 999 times. For each re-sample, the model selected for inference will be refit. This will therefore lead to new parameter estimates, and hence, corresponding different predictions for each of the groups sizes one needs to predict.

```
set.seed(12397)
B <- 999
res <- numeric(B)
gruposPredBoot <- matrix(NA,nrow=nrow(dados.filtered),ncol=999)

tableboot <- dados.filtered
tab99 <- matrix(nrow=8271,ncol=999)
tableboot <- cbind(tableboot, tab99)

for(i in 1:B){
  index = sample(1:51,51,replace=TRUE)
  dados4boot = d4reg[index,]
  mp7boot = vglm(formula = cs ~ crate, family = pospoisson, data = dados4boot)
  preds = predict(mp7boot,dados.filtered, type="response")
  tableboot[,i+19] = preds
}

#Obtaining the cs mean for each day (mean within boostraps)
tablebootday <- matrix(nrow=109,ncol=999)
B <- 999
for (i in 20:(20+B-1)){
#Cluster size mean for each day
tablebootday[,i-19] = tapply(X=tableboot[,i],INDEX=tableboot$day,FUN=mean)
}

#Density bootstrap
rdboot <- rnorm(999,0.36,0.04)
```

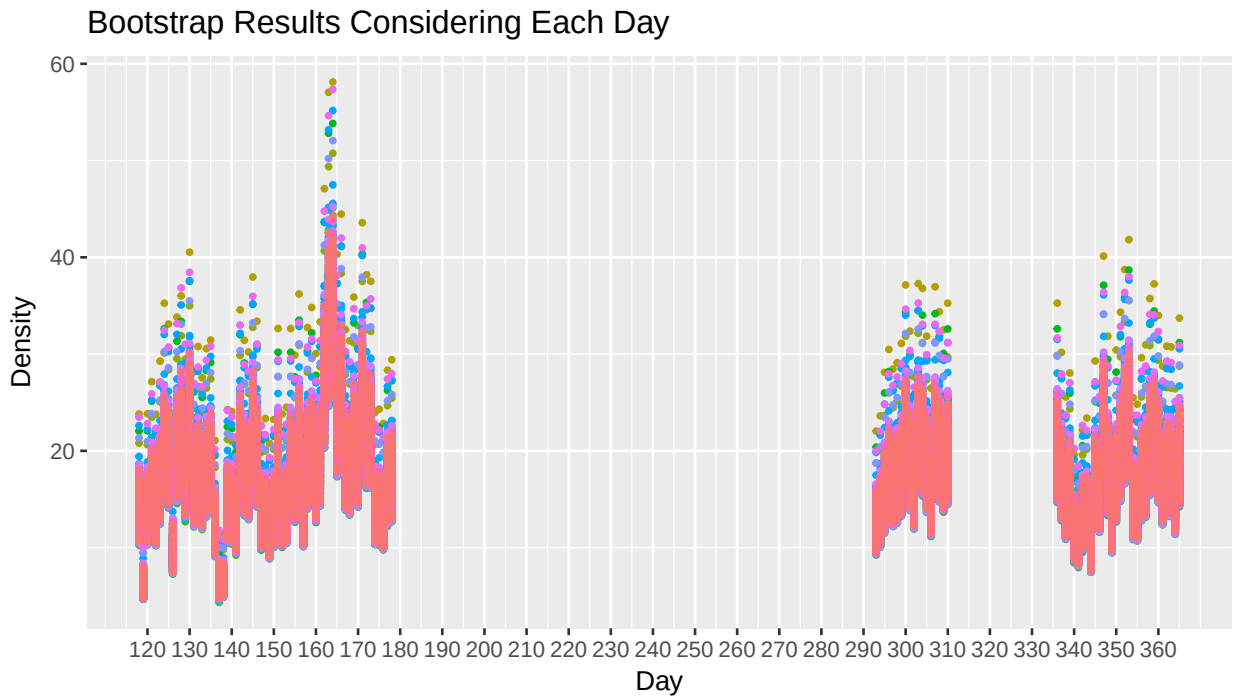
```
tableboottot <- matrix(nrow=109,ncol=999)
tablebootfix <- as.data.frame(tablebootday)
ttime <- dados.eachday$ttime
ngroups <- dados.eachday$groups
tablebootday <- rbind(tablebootday,rdboot)

#Function which calculates density for each cell
fdens <- function(x,y) #x=column, y=line
{ (((ngroups[y])*tablebootday[y,][x])/((tablebootday[110,][x])*ttime[y]))/1291*1000
}
for (i in 1:999){
  for (j in 1:109){
    tableboottot[j,][i] = fdens(i,j) }
  }

tableboottot = as.data.frame(tableboottot,header=F)
```

Boostrap plots:

```
#Click rate
fcrate <- dados.eachday$crate
tablebootdens <- tableboottot
tablebootdens <- cbind(tablebootdens,fcrate)
tablebootcs <- tablebootfix
tablebootcs <- cbind(tablebootcs,fcrate)

meltcs <- melt(tablebootcs,id="fcrate")
ggplot(meltcs,aes(x=fcrate,y=value,colour=variable,group=variable))+
  geom_point(size=0.5) + theme(legend.position="none") +
  ggtitle("Bootstrap Results Considering Each Click Rate value")
```



```r
#Density
jdays109 <- as.numeric(unique(dados.filtered$day))

daydens <-tablebootdens[-1000]
daydens <- cbind(daydens,jdays109)
meltdaydens <- melt(daydens,id="jdays109")

ggplot(meltdaydens,aes(x = jdays109,y = value,colour = variable,group = variable)) +
  geom_point(size=0.8) + theme(legend.position="none") +
  xlab("Day") +  ylab ("Density") +
  scale_x_continuous(breaks=c(120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230,
                              240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340,
                              350, 360))+ggtitle("Bootstrap Results Considering Each Day")
```



5 Conclusions

It may be concluded, based on the acoustic footprint of groups detected on AUTEC hydrophones, that the variable **click rate** appears to be the best descriptor of group size. However, when it comes to modelling, it is noticeable that more observations may be needed, as a small data set will never allow a complex model to be a parsimonious choice. Therefore, it is possible that with additional data more complex models might prove useful to describe group size from the group's acoustical footprint. Although the model composed solely by the **click rate** variable was always among the models' top 3, the variable **number of hydrophones** was replaced by **whiskey hydrophones** on the remaining two models when only considering the groups with a confidence level of 1. This may indicate difficulties when choosing between variables. Nevertheless, the model **mp7** appears to reasonably describe the response variable. In the future, more data shall be added to the modelling dataset.